1-1-2001

Single-kernel maize analysis by near-infrared hyperspectral imaging

Robert Patrick Cogdill
Iowa State University

Single-kernel maize analysis by near-infrared hyperspectral imaging

by

Robert Patrick Cogdill

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Agricultural Engineering

Major Professor: Charles R. Hurburgh, Jr.

Iowa State University

Ames, Iowa

2001

Graduate College
Iowa State University

This is to certify that the Master's thesis of

Robert Patrick Cogdill

has met the thesis requirements of Iowa State University

Signatures have been redacted for privacy

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GENERAL INTRODUCTION

As production agriculture has pushed further and further into the age of biotechnology and high-performance genetics, efficient and effective seed research and development has grown continually more vital. Along with a concentration of research effort has come an increased need for more powerful analytical tools to assist researchers in developing and understanding new organisms.

For over thirty years, near-infrared spectroscopy (NIRS) has risen to meet many of the analytical requirements presented by the maize seed breeding industry. Undoubtedly, rapid whole-grain analysis by NIRS played a vital role in the successful introduction of high-oil maize. The demand for more "designer grains" with traits such as extremely high nutrient density, specific nutrition profiles, or better environmental functionality (i.e., low phytate maize) underscores the growing pool of analytical problems that might be solved with the help of NIRS.

In light of the demands to be placed on NIRS in the future, the introduction of low-cost, tunable optical filters and high-speed digital imaging electronics is bringing about a quantum leap in the feasibility of solving such demands via NIR hyperspectral imaging. NIR hyperspectral imaging adds the ability to capture and observe the spatial variability in composition exhibited by biomaterials. By nondestructively analyzing single seeds of maize using NIR hyperspectral imaging, maize seed breeders may someday be able to more effectively quantify and track the changes in seed structure and function introduced across subsequent generations in a line of seed being developed for commerce. Or, NIR hyperspectral imaging might be a vital tool for establishing the substantial equivalency of a new organism by allowing the visualization of the spatial distribution of spectral differences between two organisms.

Before NIR hyperspectral imaging is to become a useful tool for seed analysis, techniques for gathering, handling, and analyzing the voluminous data need to be developed. Moreover, a rigorous exploration of the feasibility of NIR hyperspectral imaging of single kernels of maize for quality analysis needs to be undertaken.

Thesis Organization

This thesis, "Single-Kernel Maize Analysis by Near-Infrared Hyperspectral Imaging," is in the alternate thesis format. The paper is written in the format required for publication in the Transactions of the American Society of Agricultural Engineers (ASAE). The paper will be submitted for publication in August, 2001. General conclusions follow the paper.

Charles R. Hurburgh, Jr., co-author and major professor, and committee members Carl J. Bern, Roger Jones, and Kenneth Koehler offered their guidance and assisted in the successful completion and analysis of the research described in this paper.

SINGLE-KERNEL MAIZE ANALYSIS BY NEAR-INFRARED HYPERSPECTRAL IMAGING

A paper to be submitted for publication in the Transactions of the American Society of Agricultural Engineers (ASAE)

Robert P. Cogdill[1,2], Charles R. Hurburgh, Jr.[1], Glen R. Rippke[3], Roger W. Jones[4]

ABSTRACT

The objectives of this research were to develop a technique for creating calibrations for the prediction of the constituent concentration of single kernels of maize from NIR hyperspectral image data and to evaluate the feasibility of a NIR hyperspectral imaging spectrometer as a tool for the quality analysis of single kernels of maize. Two sets of single-kernel maize samples were analyzed by hyperspectral transmittance imaging in the range of 750 – 1090 nm. The spectral image data were standardized and standard normal variate, multiplicative scatter correction, and no preprocessing were compared before calibration spectra were extracted. Extracted spectral data were transformed by both log(1 / T) and power transformations before PLS calibration models for the prediction of moisture and oil concentration were created. Variable selection by genetic algorithm was tested as a means of reducing the amount of spectral image data needed for creating accurate calibrations. The best method for creating calibrations was inferred by maximizing the performance of the moisture calibration, for which a larger amount of more reliable reference chemistry data was available. The moisture calibration correlation and ratio of data range to standard error of cross validation (SECV) indicate that NIR hyperspectral imaging can be an effective tool for non-destructively determining the quality of single-kernels of maize. The performance and subsequent analysis of the performance of the oil calibration reveal the need for improved methods of destructive analysis for single-kernel calibration reference.

[1]Graduate student and Professor, respectively, Agricultural and Biosystems Engineering Department, Iowa State University
[2]Primary researcher and author
[3]Grain Quality Laboratory, Iowa State University
[4]Associate Scientist, United States Department of Energy, Ames Laboratory

INTRODUCTION

Near-Infrared Analysis of Biomaterials

In response to the growing demand for high-performance biomaterials analysis capabilities, near-infrared spectroscopy (NIRS) has quickly evolved from an exotic laboratory technique into a mainstay tool for a variety of qualitative and quantitative analysis tasks. The rise in the popularity of NIRS (as a method of biomaterials analysis) is due to its combination of powerful analytical capabilities, simplicity of use, and cost effectiveness.

The analytical capabilities of NIRS are rooted in the broad and repeating absorption bands exhibited by carbon-hydrogen, oxygen-hydrogen, and nitrogen-hydrogen bonds in the NIR region of the electromagnetic spectrum. While the overlapping tendencies of these absorption bands make direct interpretation of NIR absorbance spectra difficult, chemometric techniques, such as partial least squares regression (PLS), take advantage of the broad, repeating structure of the absorbance bands to yield accurate, robust calibration equations from which many constituents and quality attributes may be predicted simultaneously.

Biomaterials analysis by NIRS is generally a very simple task because little or no sample preparation is required to obtain useful measurements. In fact, NIRS was first used for analysis, as opposed to mid-infrared (mid-IR), because the relatively low absorptivities of the absorption bands are more compatible with moderately concentrated samples and longer path lengths. These longer path lengths allow spectra to be measured by transmission through intact materials, which often negates the need for sample preparation (Osborne et al., 1993). In addition to the cost savings in sample preparation, NIRS analysis is usually more cost effective than competing methodologies because the equipment is often cheaper and more readily available. NIRS equipment is also used in other, high-volume industries such as optical networking and consumer digital imaging products.

Among others, whole grain analyzers have been successfully calibrated for the reliable prediction of such constituents as moisture, crude protein, starch, fiber, and oil content (Hardy et al., 1996). An NIR bulk-grain analyzer has even been shown to be a feasible predictor of the Roundup Ready® status of whole soybeans (Roussel et al., 2001).

As a result of the quickly growing demand for higher-quality and nutrient-dense grains and oilseeds, NIRS analysis has also been adopted by the seed breeding industry as a research and development tool.

Single-Seed NIR Analysis

Stemming from the logic that only from high-quality seed will high-quality grain be grown, NIRS has been used as a segregation tool, via bulk-grain analyzers, to separate lots of breeder's seed into groups of differing quality. The higher-quality groups are planted, while the lower-quality groups of seeds are discarded, to speed the introduction or enhancement of valuable quality traits. Intuitively, the ability to analyze smaller samples of grain will yield greater increases in quality per generation where selection is employed, because fewer low-quality seeds will be planted in each group. Silvela et al. (1989), demonstrated that the rate of oil content gain was significantly greater if breeding selection occurred on a single kernel basis, as opposed to using oil values from composite samples containing all kernels on an ear.

Non-destructive single-seed quality analysis has long been a desire of seed breeders. Single-seed quality analysis via NIRS has been applied to predict oil and protein content in maize, wheat, and soybeans (Orman & Schumann, 1992; Dyer & Feng, 1995; Abe et al., 1995), moisture content in maize, lima beans, peanuts, soybeans, and sunflower (Lamb et al., 1991; Finney & Norris, 1978; Norris, 1983; Norris & Hart, 1965), oil content in meadowfoam (Patrick & Jolliff, 1997), oleic and linoleic acid in sunflower (Sato et al., 1995; Velasco et al., 1999a), oil, protein, and glucosinolate content, and oleic, linoleic, erucic acid concentration in rapeseed (Sato et al., 1998; Velasco et al., 1999b; Velasco et al., 1999c). Commercial single-seed NIRS analysis products have been offered in recent years.

While encouraging results have been reported in the prior works with single-seed NIRS analysis using traditional single-spectrum (non-imaging) machines, some seed-orientation dependent variability in the analyses was noted (Abe et al., 1995; Orman & Schumann, 1992). Moreover, some maize breeders have mentioned unacceptably high errors in the repeatability of single-kernel NIRS analyses procured via single-spectrum analyzers. They have theorized this is a result of the analyzer only 'seeing' a small portion of the maize

kernels, which are very heterogeneous in composition. Thus, prediction of the kernel's quality would not be representative of the entire kernel, but rather of a portion of the kernel. Hyperspectral NIR imaging may be a more suitable technique for quantifying the spectral characteristics of spatially heterogeneous biomaterials (such as single kernels of maize).

Hyperspectral NIRS Imaging

Hyperspectral NIRS imaging is a form of NIRS imaging where the analysis of an object includes images captured at many wavelength bands in the NIR region of the electromagnetic spectrum. Though the definition varies among researchers, hyperspectral NIRS imaging is simply an extension of multispectral NIRS imaging, where images are captured at a much smaller number of wavelength bands (usually two or three).

As image data are collected by a hyperspectral imaging system, they are often arranged into a three-way array of data, called a hypercube. The first two axes of the array contain the vertical and horizontal pixel coordinates of the images, while the third axis contains the spectral data for each pixel location within the field of view (figure 1). Thus, for example, a hyperspectral imaging system with a 512 x 512 pixel detector array would collect 262,144 spectra during each analysis. Data collected would be arranged into a 512 x 512 x m array, where m is the number of wavelengths being imaged.

While much of the spectral information is redundant, due to the high correlation between neighboring pixels, the enhanced capability of a hyperspectral imaging spectrometer to collect spatially varying spectral data is obvious. The increased capabilities (of hyperspectral NIRS imaging) come at the expense of significantly increased scan and computation times. For practical application of hyperspectral NIRS imaging, it is important to establish reasonable levels of spatial and spectral resolution as a trade-off between accuracy and computing time.

Hyperspectral NIRS imaging has been widely utilized for remote sensing and microscopy. While NIRS imaging has not been as prevalent in agricultural biomaterials analysis, it is becoming an increasingly popular methodology. Taylor & McClure (1989) pioneered NIRS imaging biomaterials analysis with their research using NIRS imaging to visualize plant stress in tobacco leaves. Evans et al. (1998), reported the construction of a

hyperspectral NIRS imaging system for the quantification of nitrogen stress on growing green bean plants.

Application of NIRS imaging of agricultural biomaterials has been for quality control and defect detection in apples (Upchurch et al., 1994; Bellon-Maurel et al., 1992), peaches and apricots (Miller & Delwiche, 1990; Zwiggelaar et al., 1996), beef (Hatem et al., 1999), and poultry carcasses (Park et al., 1998). Sugiyama (1999) constructed and calibrated a multispectral imaging system to predict the distribution of sugar in the cross-section of ripe melons. Ridgeway & Chambers (1998) used NIRS imaging to nondestructively detect insects inside single-kernels of wheat. Archibald et al. (1998) developed a system to analyze wheat and predict the color classification on a single-kernel basis.

OBJECTIVES

The objectives of this work were to:

1. Develop a technique for creating calibrations for the prediction of the constituent contents of single kernels of maize from NIR hyperspectral image data.
2. Evaluate the feasibility of a NIR hyperspectral imaging spectrometer as a tool for the quality analysis of single kernels of maize.

MATERIALS AND METHODS

Imaging Equipment

The imaging equipment consisted of a detector, tunable optical filter, sample stage, collimating optics, and light source (figure 2). The detector was an Apogee KX-260 monochrome scientific camera, which utilized a thermoelectrically cooled, 512x512 pixel, silicon CCD array. The imaged data was digitized to a quantization level of 14 bits by an on-board microprocessor (no frame grabber is necessary with the KX-260).

The tunable optical filter was a CRI Varispec® model VS-NIR liquid crystal tunable filter (LCTF) with the following specifications: 55 mm clear aperture, 700-1100 nm tunable range, 10 nm bandwidth. A LCTF was chosen as the method of wavelength selection due to its combination of high tuning speed, appropriate range, narrow bandpass, and ease of use.

The LCTF is becoming a popular tool for hyperspectral NIRS imaging research (Archibald et al., 1998; Evans et al., 1998; Miller & Hoyt, 1995).

The light source, a 250 Watt, 24 volt tungsten-halogen lamp, was powered by a DC source to minimize 60 Hz noise. The collimating optics consisted of a 700 nm long-pass filter (to reduce the effects of out-of-band light), a focusing lens, a diffuser, and a flat mirror which directed the light onto the underside of the sample presentation stage. The sample presentation stage consisted of a silica glass microscope slide in a stable frame. The collimating optics were adjusted to maximize uniformity of illumination by varying the positions of the focusing lens and the diffuser.

Image Acquisition and Analysis Software

The imaging equipment was controlled by custom-written software (Liu, 2001). The software allowed the user to collect hyperspectral image data by controlling the LCTF and camera simultaneously. As data were collected it was stored on the hard drive of a Dell Dimension XPS T450 Pentium II PC. Further processing of the image data and the chemometric modeling were implemented using the Image Processing Toolbox (The Mathworks Inc.), the PLS_Toolbox (Eigenvector Research Inc.), and a variety of custom-written MATLAB functions.

Calibration Sample Set

The moisture and oil calibration sample sets were selected based on constituent composition and visual uniqueness in an attempt to span as much of the sample spectral space as possible. All of the kernels in the oil calibration set were chosen from populations represented by kernels in the moisture calibration set. The moisture and oil calibration data sets are described in table 1. The single-kernel moisture reference data were obtained gravimetrically. Single-kernel samples were placed in a forced-air oven at 103°C for 72 h, and moisture was computed on a wet basis according to ASAE method S352.2 DEC92.

Oil reference chemistry data were obtained by performing supercritical fluid extraction (SFE) on the single kernel samples, via a LECO FastFat™ HT supercritical fluid extractor. Prior to oil extraction, the single kernel samples were crushed and ground using a

mortar and pestle. The crushed kernel was then weighed and subjected to SFE. The extracted oil was weighed and the percentage oil content of the crushed kernel was then determined gravimetrically according to AOCS official method Am 3-96.

The significant difference in calibration set size between the moisture and oil calibration sets reflects the greater time and labor involved with the supercritical fluid extraction method (as compared to the oven moisture determination).

Hyperspectral Image Acquisition and Preprocessing

To form the hyperspectral image set for each sample, the kernels were placed on the sample stage at random positions in the field of view of the spectrometer. The results of a preliminary experiment indicated that kernel positioning did not have any significant effect on the kernel analysis. Images of each sample were taken at 5 nm intervals, covering the range from 750 to 1090 nm. The exposure time was set differently for each wavelength imaged in an attempt to compensate for the detector sensitivity, which varied across the spectral range. The exposure time was set to maintain an average intensity of 10,000 A/D counts (out of a possible 16,383 for a 14-bit camera) within a 200x200-pixel region in the center of an image of a piece of opal glass being illuminated on the sample stage.

To complete the hyperspectral image set of the sample, two images were taken to construct a binary mask (figure 3) of the maize kernel; an image was taken of the sample with a piece of double-ground quartz diffuser material under the sample stage, and an image was taken of the double-ground quartz with the sample removed. The absolute value of the arithmetic difference between these two images was thresholded such that pixels in positions where there was no kernel were set to logical zero, and pixels in positions where there was a portion of the kernel were set to logical one. The binary mask was then used to omit the data from pixels outside the boundaries of the sample. The remaining pixel data were rearranged into a two-way array (Geladi & Grahn, 1996) where each row of the array contained the location and spectral information for every pixel that remained after masking. The result is an N x 71 array, where N is the number of pixels that were equal to logical 1 in the binary mask and thus contained relevant spectral data.

Along with the hyperspectral image set for each sample, a standardization image set was acquired once during every data collection session. The purpose of the standardization image set was to compensate for temporal differences in the spectrometer's response both across the field of view and across the spectral range. The standardization image set consisted of dark reference, light reference, and empty sample stage images. Dark reference images were acquired by reading the detector array with the lens cap on. The dark reference is an image of the inherent additive noise of the CCD array plus whatever thermal radiation reaches the detector. The light references were images of a piece of opal glass. The empty sample stage images were images of the sample stage with the light source powered, but no sample in the field of view. The empty sample stage images were subtracted from the sample images to remove the baseline effects of the apparatus. Intuitively, this would be similar to subtracting the spectrum of an empty sample holder from sample spectra collected using a single-spectrum spectrometer. Both the light reference and empty sample stage images were taken at the same wavelength and exposure times as the images of the samples. The standardization image set was applied to all N pixels in the hypercube according to the following formula:

$$I_{\lambda,n} = \left[C\left(S_{\lambda,n} - D_n\right) \div \left(L_{\lambda,n} - D_n\right)\right] - \left[C\left(E_{\lambda,n} - D_n\right) \div \left(L_{\lambda,n} - D_n\right)\right] \quad (1)$$

$I_{\lambda,n}$ = standardized image pixel n, at wavelength λ

$S_{\lambda,n}$ = sample image pixel n, at wavelength λ

$E_{\lambda,n}$ = empty sample stage image pixel n, at wavelength λ

$L_{\lambda,n}$= light reference image pixel n, at wavelength λ

D_n = dark reference image pixel n

C = mean brightness level within a 100x100-pixel region in the center of $(L_{850,(1:N)} - D_{(1:N)})$

Besides standardizing the hypercube, some additional preprocessing methods were tested. In light of the variability in particle size and pathlength that is encountered by photons traveling through the kernel, spectral preprocessing should be beneficial. The performance of models implementing standard normal variate (SNV) preprocessing and models implementing multiplicative scatter correction (MSC) were compared to models developed without preprocessing. All preprocessing models used the moisture calibration dataset.

SNV, which can be thought of as mean centering and scaling in the object direction, was applied using the procedure described by Barnes et al. 1989. Outlier detection was performed among the spectra within the hypercube, to reject spectra with a standard deviation near zero (across wavelengths). Spectra with exceedingly high standard deviation was exceedingly high were also excluded, since they are likely to contain false data. The within-hypercube outlier rejection was applied to all data prior to chemometric modeling.

MSC was applied using the PLS_Toolbox function MSCORR, according to the method outlined in Martens & Naes (1989). Each sample spectrum was offset and rotated according to the inverse of a linear fit between a block of sample data and a reference spectrum. The reference spectrum used for every sample was the global mean spectrum from the moisture and oil calibration set.

Spectra Extraction and Calibration Development

Following the hypercube standardization and preprocessing a single spectrum for each sample was extracted (for chemometric modeling) by taking the arithmetic mean of the N standardized intensities at each wavelength, the result being a 1x69 array of optical transmittance ratios. Thus, every hypercube (kernel) would be represented by a single extracted spectrum during calibration and prediction.

Spectral data was converted from transmittance ratio to optical absorbance units. Transmittance NIR spectrometric absorbance data are generally reported as the log of their inverse, according to the Beer-Lambert equation; when pathlength and molar absorptivity are constant, the NIR absorbance will vary linearly with the concentration of the analyte in the matrix, where NIR absorbance is the log of the inverse of the optical transmittance of the sample. While this equation works well for modeling concentration by transmittance through transparent liquids and many bulk particulate samples, it may be less than optimal for describing transmittance through a heterogeneous, solid matrix. The purpose of transforming data is to improve the effectiveness of linear modeling, which depends on an assumption of normally distributed data. The transformation that best fits the inverse of the specific non-linearity of the data will be optimal for linear modeling.

A variety of power transformations (which were near log(1 / T)) were evaluated for their effect on calibration performance. Specifically, the spectral data were transformed by:

$$Z = X^P \qquad (2)$$

Z = transformed spectral data

X = un-transformed spectral data

P = power transformation constants chosen to cover the range [-2,2] in increments of 0.1

For all calibration data sets, the spectral data and reference chemistry data were compiled and scaled to zero-mean and unity standard deviation.

Before preprocessing and data transformation model comparisons were performed, between-hypercube outlying samples were identified and removed according to the irregularity of their extracted spectra and prediction residual. Sample spectra whose Hotellings T^2 statistic exceeded the 99% multivariate confidence interval were omitted from further analysis. The T^2 statistic and confidence interval were calculated from the principal component scores for each spectrum using the pca.m function in the PLS_Toolbox; 20 latent variables were used for the principal components decomposition. The moisture and oil calibration spectra sets were combined for spectral outlier detection.

For both the moisture and oil calibration, when the first calibrations were tested, samples whose prediction residual exceeded the 99% student's-t confidence interval were identified as outliers and removed from further modeling. Outliers were removed only once for the moisture and oil calibrations.

Exploratory models of the data were inferred by partial least squares (PLS) regression. Preprocessing methods and model performance were evaluated by using k-block cross-validation. Model performance is the average standard error of cross validation (SECV) with k set to 10. The SECV is calculated with the following formula:

$$\mathrm{SECV} = \left[\left(\sum_N \left(Y_p - Y_A\right)^2\right) \div n\right]^{0.5} \qquad (3)$$

Y_P = predicted value from model output

Y_A = actual value from model reference data

n = number of samples in the data set

Variable Selection by Genetic Algorithm

Once the standardization and data pretreatment methods had been set by selecting the methods that minimized SECV, a genetic algorithm search was employed to find an optimal subset of the original 69 wavelengths for future analyses. The motivation behind the genetic algorithm search was to reduce the amount of data being handled and to omit the noise added to the model by extraneous spectral data. The genetic algorithm (GA) is one of a family of adaptive search methods that is modeled after the genetic evolutionary process. An attractive feature of the GA is its efficiency when dealing with difficult combinatorial search problems; genetic algorithms have the characteristic ability to avoid local extrema because of their parallel exploration of the search space (Goldberg, 1989a; Tang, 2000).

The genetic algorithm provided in the PLS_Toolbox was used for variable selection. Implementation of the GA required setting a number of parameters: population size, window width, % initial terms, max generations, % at convergence, mutation rate, and the crossover setting. The population size was set to be approximately equal to the chromosome length (Goldberg, 1989b), which in our case was 69; the nearest setting available with the GA implemented was 68. The window width was set to one, and the % initial terms was set at 30% (PLS_Toolbox default). The max generations was set to 225, though experience has shown that the algorithm will usually converge much sooner. The % at convergence parameter, which stops the algorithm when that percentage of chromosomes in the current population have the same solution, was set very high (70%). While this slows the convergence of the GA, it can aid in the interpretation of the results by presenting a smaller population of solutions. The mutation rate was set to 0.007, which is inversely proportional to the population size (Goldberg, 1989a). The crossover parameter was set to double crossover. PLS regression with a maximum of 25 latent variables, and MLR were used as fitness functions. Cross-validation was set to random, with 10 subsets and 1 iteration.

RESULTS AND DISCUSSION

Preprocessing Methods

The results of applying SNV, MSC, and no spectral preprocessing are in table 2. SNV outperformed MSC and no preprocessing when no other transformation was applied. However, after further data transformations were applied to the calibration data, SNV and MSC preprocessing were detrimental to calibration performance. MSC removed data essential to calibration performance. However, the results of SNV versus no preprocessing are more intriguing. One explanation may be that SNV preprocessing removes information that is only useful to a linear model after further data transformations have been applied. For all further calibrations, no preprocessing was implemented.

Data Transformation

The performance of the various data transformation methods is given in figure 4. While the log(1/T) transformation improved model performance versus un-transformed data, SECV = 1.36; the P = 0.4 power transformation consistently performed the best of all transformations tested. The simplest explanation for the relative superiority of the power transformation is that the shape of the power transformation is a better approximation of the inverse of the non-linearity observed during NIR transmittance through single kernels of maize (figure 5). Since the results were established using random cross-validation, it is reasonable to assume that the results are sufficiently general to apply to future data sets acquired by the same method. However, some data set dependence was observed, which suggests that some other global optimum transform is likely to exist. All further calibration models were produced with spectral data that was transformed according to (2), with P = 0.4.

Initial Calibration Results

The performance of the moisture calibration (after all preprocessing and transformation, but before variable selection) is summarized in figure 6. The observed range:SECV value of 16.8 and R^2 of 0.847 indicate that the calibration would be effective for quantitative moisture content prediction. According to the AACC Approved Method 39-00, a

ratio of range to SECV between four and eight indicates the possibility of distinguishing between high and low values; with a ratio of range to SECV between eight and twelve, there is a possibility of making quantitative predictions; a ratio of range to SECV greater than twelve indicates good predictability of constituents. The observed calibration performance leaves significant room for improvement, however, when compared to the correlation often achieved during bulk-sample maize moisture analysis by single-spectrum analyzers. The observed calibration coefficient vector (figure 7) is much more jagged than is usually seen with instruments working in a comparable range (figure 8), suggesting increased noise in the spectral data and reference chemistry is limiting the calibration performance. However, as more data are added to the calibration set, it is likely that the difference in performance will diminish.

The performance of the initial oil calibration is summarized in figure 9. While the range:SECV of 7.9 and R^2 of 0.469 indicate that the oil calibration may be useful for segregating high and low oil concentration kernels, its performance was significantly lower than was observed with the moisture calibration. While the oil calibration coefficient vector is also jagged (figure 10), the relative magnitude of the largest coefficients is much smaller than that of the moisture calibration coefficient vector. The magnitude of the vectors can be compared because both the reference chemistry and the spectral data have been scaled to zero mean and unity standard deviation. Furthermore, an abnormal amount of over-fitting occurs with the inclusion of an excess of PLS factors in the calibration, as can be seen in the scree plot for the oil calibration cross-validation (figure 11). This phenomenon generally occurs during PLS model development using cross-validation when the noise level in the calibration data set is high. Considering the extreme difference in performance between the moisture and oil calibrations, it was assumed that much greater error in the oil reference chemistry was the root of the problem.

Repeated Measures Test

In order to better interpret the performance of the oil calibration, a set of experiments was performed to estimate the contributions to the model error made by the reference chemistry, the spectrometer, and the calibration, according to the following model:

$MSE_{total} = [MSE_1 + MSE_2] + [MSE_3 + MSE_4]$ (% oil concentration) (4)

MSE_{total} = total observed mean squared error of the model ($SECV^2$) ≅ 1.46 %

$[MSE_1 + MSE_2]$ = mean squared error of the reference chemistry repeatability ≅ 0.72 %

MSE_3 = mean squared error of the spectrometer repeatability ≅ 0.27 %

MSE_4 = mean squared calibration error

The goal of the repeated measures test was to estimate MSE_1, MSE_2, and MSE_3 to solve for MSE_4 by satisfying equation (4).A truer estimate of the predictive ability of the spectrometer would be embodied in $[MSE_3 + MSE_4]$.

The decision to split the reference chemistry error into two parts (MSE_1 and MSE_2) was made because the reference chemistry was destructive. Oil can be extracted from a single maize kernel only once. Lots of ground grain samples (for which heterogeneity of composition can be assumed) were used to estimate the repeatability of the SCFE reference method. Using the same method as for the single kernel oil reference chemistry (sans crushing and grinding by mortar and pestle), the percent oil concentration was determined for eight samples of approximately 0.3 grams drawn from each of five lots of ground maize. The mean of the within-group variance of the five lots of eight samples was the estimate of MSE_1. The purpose of MSE_2 is to account for the error under-estimation due to the difference between extracting the oil from single kernels that have been pulverized with a mortar and pestle and machine-ground samples of bulk grain. Because it is impossible to estimate MSE_2 with confidence, it is assumed to be an unknown value greater than or equal to zero, changing (4) to an inequality.

To estimate MSE_3, three repetitions of spectral data and oil content predictions were obtained from twenty-eight kernels of maize. The mean of the within-group variance of the predicted oil concentration for the twenty-eight lots of three samples was the estimate of MSE_3. To obtain an estimate of the effect of kernel positioning on the sample prediction, fourteen of the twenty eight kernels were analyzed three times without being repositioned in the field of view (MSE_{3a}) and fourteen kernels were analyzed in a different position for each repetition (MSE_{3b}). For the first repetition, the kernels were placed randomly in the field of view, for the second repetition, the kernel was rotated 90°, and for the third repetition, the kernel was flipped. MSE_3 was then the result after pooling the data for MSE_{3a} and MSE_{3b}.

By solving (4) for MSE_4 ($\cong$ 0.47 %) it becomes clear that error in the reference chemistry is to blame for (at least) half of the model error. This would suggest that further work towards improving the oil calibration should focus on reducing the reference chemistry error. Drastically increasing the size of the oil calibration spectral database would also improve the performance of the calibration, as the impact of random error would diminish (according to the central limit theorem).

Kernel positioning has no real effect on the single kernel analysis since the observed MSE_{3b} was actually lower than MSE_{3a} (though is not statistically significant). These results affirmed similar observations made during preliminary experiments.

Variable Selection by Genetic Algorithm

The results for the variable selection by GA are very promising. For the moisture calibration, the GA identified a subset of 29 wavelengths to maximize the moisture calibration performance of PLS and MLR models, a 58% reduction in the amount of data that must be handled by the system. The moisture calibration SECV decreased from 1.21 to 1.11%.

The wavelength bands selected by the GA also provide rationale for confidence in the performance of the spectrometer by relating the selected bands to pure-component absorbance peaks. A band of the wavelengths selected (figures 12 & 13) were near 970 nm, which corresponds to the location of an absorbance peak in the pure-component spectrum of water (Williams & Norris, 1987). Another, broader set of wavelengths was selected in the region of 880 - 940 nm, which corresponds to the location of peaks in the protein and oil pure-component spectra. While this may seem odd at first, after noticing that the regression coefficients for the two groups of wavelengths are of opposite sign (figure 14), it might be deduced that the calibration is modeling a negative correlation between protein and oil content and water.

The GA search yielded similar results for the oil calibration, identifying 20 wavelengths for PLS, and 15 wavelengths for MLR, a data reduction of nearly 79%. An even greater increase in model performance was noticed for the oil calibration, as the SECV decreased from 1.51 to 1.20 %. Moreover, it can be seen in figure 15 that over-fitting was

reduced significantly by omitting the unnecessary spectral data. As with the moisture calibration, some of the wavelengths selected for optimal oil calibration performance lend themselves to direct interpretation (figures 16 & 17). The two largest positive coefficients in the regression vector (figure 18) are at 910 and 955 nm, which roughly correspond to protein and water, respectively. The two most negative coefficients are at 850 and 925 nm, peaks in benzene and oil, respectively. The model is apparently capitalizing on the negative correlation between the major constituents of maize.

An experiment was performed to determine if the model performance observed after variable selection by genetic algorithm could be duplicated by simply omitting every other spectral variable, resulting in a 10-nm spectral resolution. As a result of the dimension omission, the full-spectrum moisture calibration SECV increased from 1.21 to 1.36% and the full-spectrum oil calibration SECV increased from 1.51 to 1.55%. These results affirm the use of a GA to select optimal variables for modeling. Moreover, these results suggest that even better model performance might be achieved by sampling at an even higher spectral sampling rate before selecting optimal variables for modeling.

CONCLUSIONS

The objectives of this study were to develop a technique for creating calibrations for the prediction of the constituent concentration of single kernels of maize from NIR hyperspectral image data and to evaluate the feasibility of a NIR hyperspectral imaging spectrometer as a tool for the quality analysis of single kernels of maize. These were met by comparing the performance of various methods of treating NIR hyperspectral transmittance image data for chemometric models to predict the moisture and oil concentration of single kernels of maize. The following conclusions can be drawn from the results of the study:

1. Single-kernel analysis by hyperspectral NIR imaging can be a useful technique for quantitative prediction of the moisture concentration of single kernels of maize, and for screening single kernels of maize for differentiating between ranges of oil content.
2. Results indicate that the log(1 / T) is not likely to be the best transformation for converting transmittance ratio spectra (collected from hyperspectral images of single kernels of maize) into absorbance spectra.

3. When combined with an appropriate transformation to absorbance spectra, preprocessing image hypercubes by SNV or MSC before extracting calibration spectra is detrimental to calibration performance compared to using no additional preprocessing.
4. The orientation of the kernel on the imaging stage does not influence the result of NIR image analysis.
5. Reducing the number of wavelengths included in the hyperspectral image set has a positive impact on the performance of the hyperspectral imaging spectrometer for predicting moisture and oil concentration in single kernels of maize.

ACKNOWLEDGEMENTS

This research was supported in-part by ExSeed Genetics, and the Institute for Physical Research and Technology.

Special thanks go to Erica Searcy for her contribution of high-quality illustrations.

REFERENCES

1. Abe, H., T. Kusama, S. Kawano and M. Iwamoto. 1995. Non-destructive determination of protein content in a single kernel of wheat and soybean by near infrared spectroscopy. In: A.M.C. Davies & P. Williams (Eds.), *Near Infrared Spectroscopy: the Future Waves*, pp. 457-461. NIR Publications, Chichester, UK.
2. Archibald, D.D., C.N. Thai and F.E. Dowell. 1998. Development of short-wavelength near-infrared spectral imaging system for grain color classification. *Proceedings of SPIE*, SPIE vol. 3543, pp. 189-198.
3. Barnes, R.J., M.S. Dhanoa and S.J. Lister. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5), pp. 772-777.
4. Bellon, V., G. Rabatel and C. Guizard. 1992. Automatic sorting of fruit: sensors for the future. *Food Control*, January, pp. 49-54.
5. Dyer, D.J., and P. Feng. 1995. Near-infrared applications in the development of genetically altered grains. In: A.M.C. Davies & P. Williams (Eds.), *Near Infrared Spectroscopy: the Future Waves*, pp. 490-493. NIR Publications, Chichester, UK.
6. Evans, M.D., C.N. Thai and C.J. Grant. 1998. Development of a spectral imaging system based on a liquid crystal tunable filter. *Transactions of the ASAE*, 41(6), pp. 1845-1852.
7. Finney, Jr., E.E., and K.H. Norris. 1978. Determination of moisture in corn kernels by near-infrared transmittance measurements. *Transactions of the ASAE,* 21(3):581.

8. Geladi, P., and H. Grahn. 1996. In: *Multivariate image analysis*. New York, NY.: John Wiley & Sons, Inc.

9. Goldberg, D.E. 1989a. In: *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, Mass.: Addison-Wesley Publishing Co., Inc.

10. Goldberg, D.E. 1989b. Sizing populations for serial and parallel genetic algorithms. In: *Proceedings of the 3rd International Conference on Genetic Algorithms*, pp. 70-79. San Mateo, Calif: Morgan Kaufmann.

11. Hardy, C.L., G. Rippke, C.R. Hurburgh, Jr., and T.J. Brumm. 1996. Calibration and field standardization of Foss Grainspec Analyzers for corn and soybeans. In: A.M.C. Davies & P. Williams (Eds.), *Near Infrared Spectroscopy: the Future Waves*, pp. 122-131. NIR Publications, Chichester, UK.

12. Hatem, I., J. Tan and P. Shatadal. 1999. Beef quality prediction by using near-infrared image features. Presented at the 1999 ASAE/CSAE-SCGR Annual International Meeting, Toronto, Ontario Canada. ASAE Paper No. 993159.

13. Liu, J. 2001. Design and implementation of real time image acquisition and processing systems. M.S. thesis. Ames: Iowa State University.

14. Lamb, D.T., and C.R. Hurburgh, Jr. 1991. Moisture determination in single soybean seeds by near-infrared transmittance, *Transactions of the ASAE*, 34(5):2123-2129.

15. Martens, H., and T. Naes. 1989. In: *Multivariate calibration*. New York, NY.: John Wiley & Sons, Inc.

16. Miller, B.K., and M.J. Delwiche. 1990. Spectral analysis of peach surface defects. ASAE Paper No. 90-6040. St. Joseph, MI.:ASAE.

17. Miller, P.J., and C.C. Hoyt. 1995. Multispectral imaging system with a liquid crystal tuneable filter. Proceedings of SPIE, SPIE vol. 2345, pp. 354-365.

18. Norris, K.H. 1983. In: *Food Research and Data Analysis*, pp. 95 – 113, edited by H. Martens and H. Rasswum, Jr., New York, NY. Applied Sciences Publishers.

19. Norris, K.H., and J.R. Hart. 1965. Direct spectrophotometric determination of moisture content of grain and seeds, In: *Principles and methods of measuring moisture in liquids and solids. Vol. 4. Proceedings of the 1963 International Symposium on Humidity and Moisture.* pp. 19-25, New York, NY. Reinhold Publishing Co.

20. Orman, B.A., and R.A. Schumann, Jr., 1992. Nondestructive single-kernel oil determination of maize by near-infrared transmission spectroscopy, *Journal of the American Oil Chemists' Society,* Vol 69, pp. 1036-1038.

21. Osborne, B.G., T. Fearn and P.H. Hindle. 1993. In: *Practical NIR spectroscopy*, 2nd edition, New York, NY. Longman Scientific and Technical, copublished with John Wiley & Sons, Inc.

22. Park, B., R. Chen and M. Nguyen. 1998. Multi-spectral image analysis using neural network algorithm for inspection of poultry carcasses, *Journal of Agricultural Engineering Research*, Vol. 69(4), pp. 351-363.

23. Patrick, B.E., and G.D. Jolliff, 1997. Nondestructive single-seed oil determination of meadowfoam by near-infrared transmission spectroscopy, In: *Journal of the American Oil Chemists' Society,* Vol. 74, pp. 273-276.

24. Ridgeway, C., and J. Chambers. 1998. Detection of insects inside wheat kernels by NIR imaging, *Journal of Near Infrared Spectroscopy*, Vol. 6, pp. 115-119.

25. Roussel, S.A., C.L. Hardy, C.R. Hurburgh, Jr., G.R. Rippke. 2001. Detection of Roundup Ready™ soybeans by near-infrared spectroscopy, *Applied Spectroscopy*, submitted in 2000.

26. Sato, T., Y. Takahata, T. Noda, T. Yanagisawa, T. Morishita and S. Sakai. 1995. Nondestructive determination of fatty acid composition of husked sunflower seeds by near-infrared spectroscopy, *Journal of the American Oil Chemists' Society,* Vol. 72, pp. 1177-1183.

27. Sato, T., I. Uezono, T. Morishita and T. Tetsuka. 1998. Nondestructive estimation of fatty acid composition in seeds of Brassica napus L. by near-infrared spectroscopy, *Journal of the American Oil Chemists' Society,* Vol. 75, pp. 1877-1881.

28. Silvela, L., R. Rodgers, A. Barrera and D.E. Alexander. 1989. Effect of selection intensity and population size on percent oil in maize, *Theoretical and Applied Genetics,* Vol. 78, pp. 298-304.

29. Sugiyama, J. 1999. Visualization of sugar content in the flesh of a melon by near-infrared imaging, *Journal of Agriculture and Food Chemistry*, Vol. 47, pp. 2715-2718.

30. Tang, L., L. Tian and B.L. Steward. 2000. Color image segmentation with genetic algorithm for in-field weed sensing, *Transactions of the ASAE,* Vol. 43(4), pp. 1019-1027.

31. Taylor, S., and W.F. McClure. 1989. NIR Imaging spectroscopy: measuring the distribution of chemical components, In: *Proceedings of the Second International Near Infrared Spectroscopy Conference*, pp. 393-404, Korin Publishing, Japan.

32. Upchurch, B.L., J.A. Throop and D.J. Aneshansley. 1994. Influence of time, bruise-type, and severity on near-infrared reflectance from apple surfaces for automatic bruise detection, *Transactions of the ASAE*, Vol. 37(5), pp. 1571-1575.

33. Velasco, L., B. Pérez-Vich and J.M. Fernández-Martinez. 1999a. Nondestructive screening for oleic and linoleic acid in single sunflower achenes by near-infrared reflectance spectroscopy. *Crop Science*, Vol. 39, pp. 219-222.

34. Velasco, L., C. Möllers and H.C. Becker. 1999b. Estimation of seed weight, oil content, and fatty acid composition in intact single seeds of rapeseed by near infrared reflectance spectroscopy. *Euphytica*, Vol.106, pp. 79-85.

35. Velasco, L., C. Möllers and H.C. Becker. 1999c. Screening for quality traits in single seeds of rapeseed by near-infrared reflectance spectroscopy. In: N. Wratten & P.A. Salisbury (Eds.) *New Horizons for an old crop, Proceedings of the 10th International Rapeseed Congress.* The Regional Institute Ltd., Canberra, Australia.

36. Williams, P., and K. Norris. 1987. In: *Near Infrared Technology in the Agricultural and Food Industries*. pp. 247 – 290. American Association of Cereal Chemists Inc.: St. Paul.

37. Zwiggelaar, R., Q. Yang, P. Garcia and C.R. Bull. 1996. Use of spectral information and machine vision for bruise detection on peaches and apricots, *Journal of Agricultural Engineering Research*, Vol. 63(4), pp. 323-331.

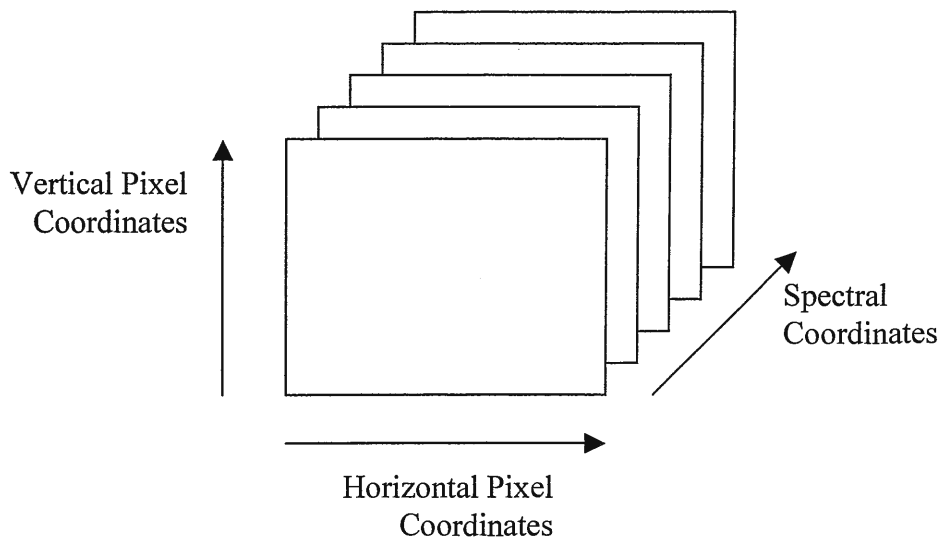

Figure 1. Illustration of the arrangement of spectral images to form a hypercube.

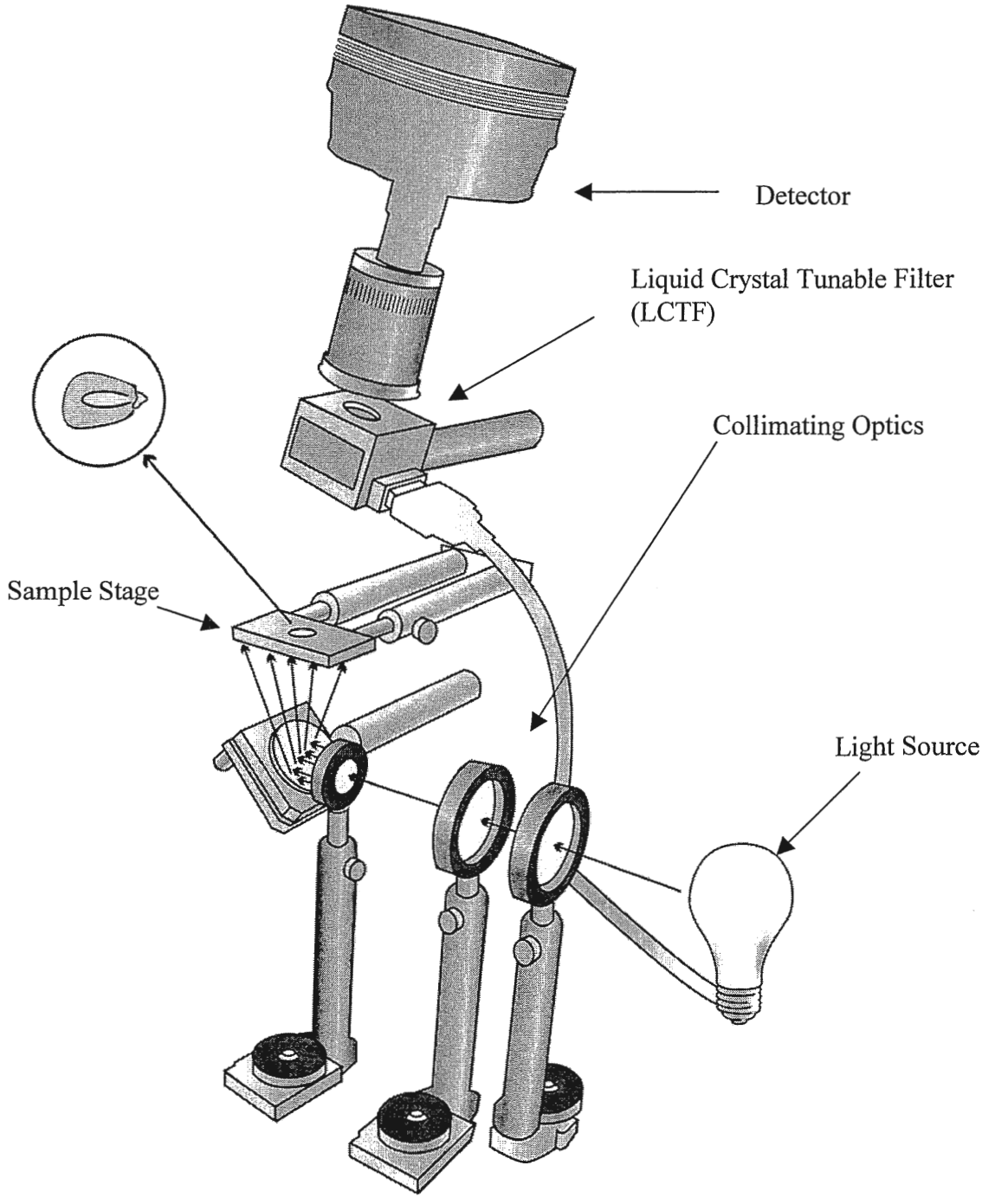

Figure 2. Illustration of hyperspectral imaging spectrometer.

Table 1. Final calibration statistics (after variable selection by genetic algorithm).

Model Basis		Calibration	
		Moisture[b]	Oil[c]
PLS	Number of Wavelengths	29	20
	Latent variables	17	15
	Correlation, r	0.933	0.806
	Correl. Coeff., r^2	0.871	0.650
	SECV	1.11	1.20
	Range:SECV	18.4	9.9
MLR	Number of Wavelengths	29	15
	Correlation, r	0.932	0.803
	Correl. Coeff., r^2	0.868	0.645
	SECV	1.16	1.22
	Range:SECV	17.5	9.7
Reference Analysis[a]			
	Number of Samples	452	153
	Average	15.66	3.19
	Maximum	30.11	12.16
	Minimum	9.74	0.26
	Spectral Outliers Removed, %	6.9	5.9
	Prediction Residual Outliers Removed, %	1.8	2.9

[a] Reference analysis statistics were calculated after outliers were removed
[b] % moisture concentration (wet basis)
[c] % oil concentration ('as is' moisture basis)

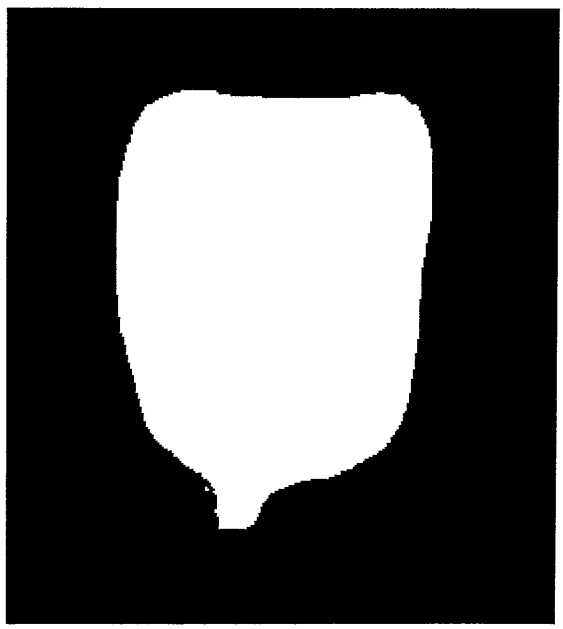

Figure 3. Typical binary mask of a single kernel of maize.

Table 2. Effect of data pretreatment methods on moisture calibration performance (before variable selection by genetic algorithm).

	Before Transformation			After Transformation			
Preprocessing Treatment	**SECV (%)**	**R2**	**lv**	**SECV (%)**	**R2**	**lv**	**Optimal Transform**
Raw Transmittance	**1.43**	**0.783**	**8**	**1.21**	**0.847**	**18**	**P = 0.4**
SNV	**1.32**	**0.817**	**16**	**1.23**	**0.840**	**20**	**P = -0.8**
MSC	**1.81**	**0.662**	**10**	**1.69**	**0.703**	**11**	**P = 0.2**

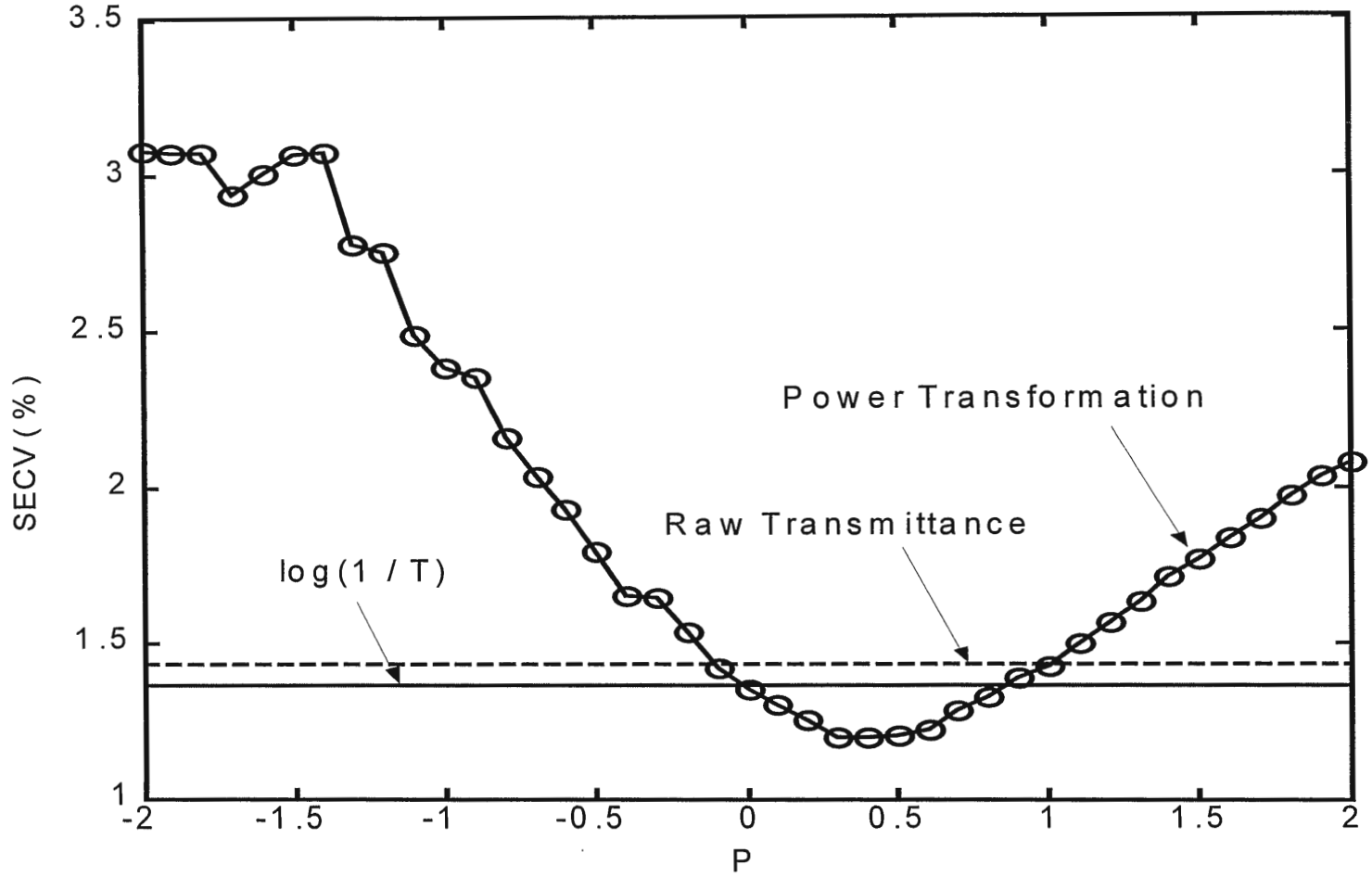

Figure 4. Effect of various data transformations on moisture calibration performance.

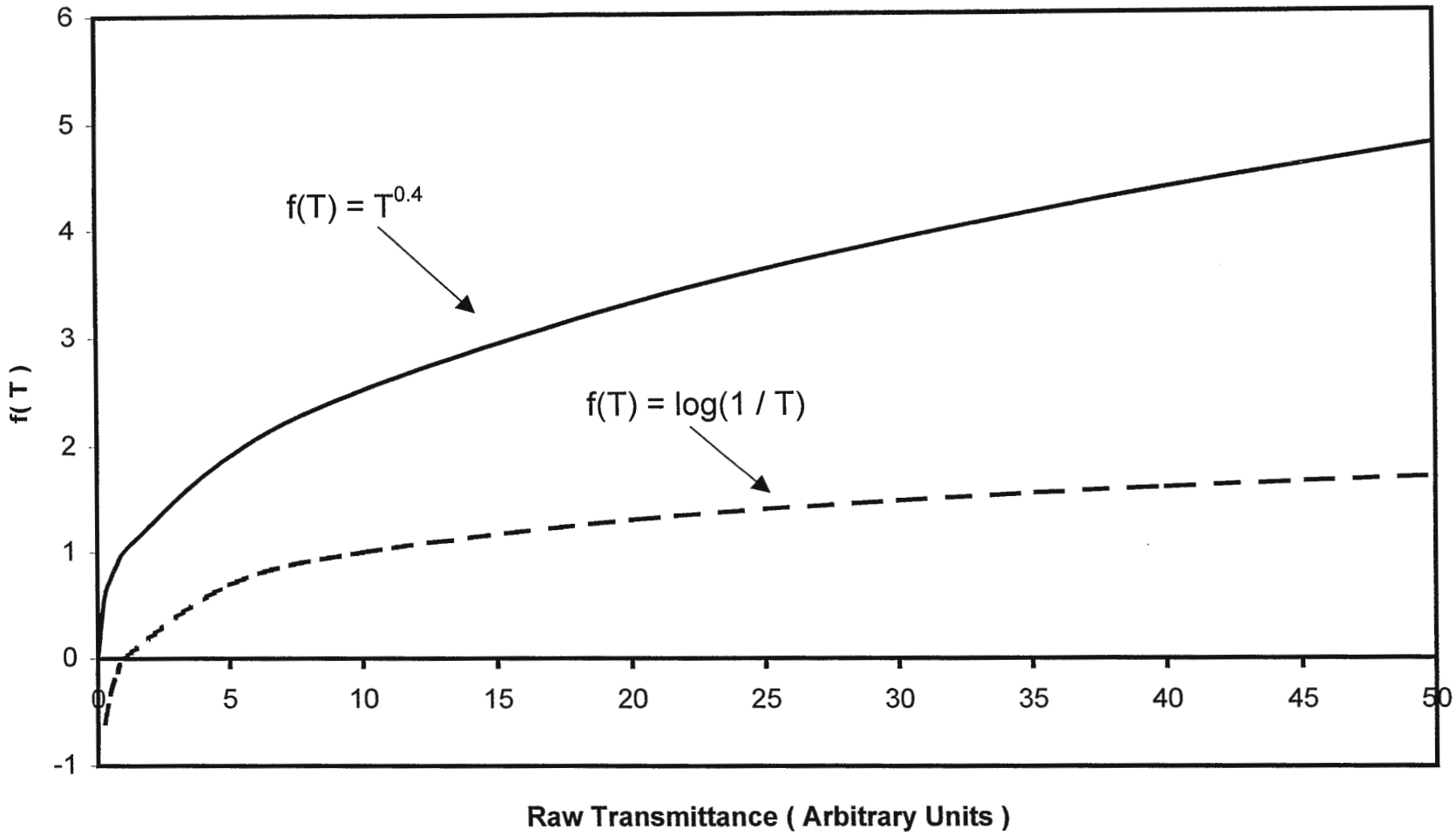

Figure 5. Illustration of the relationship between log(1 / T) and P = 0.4 power transformation.

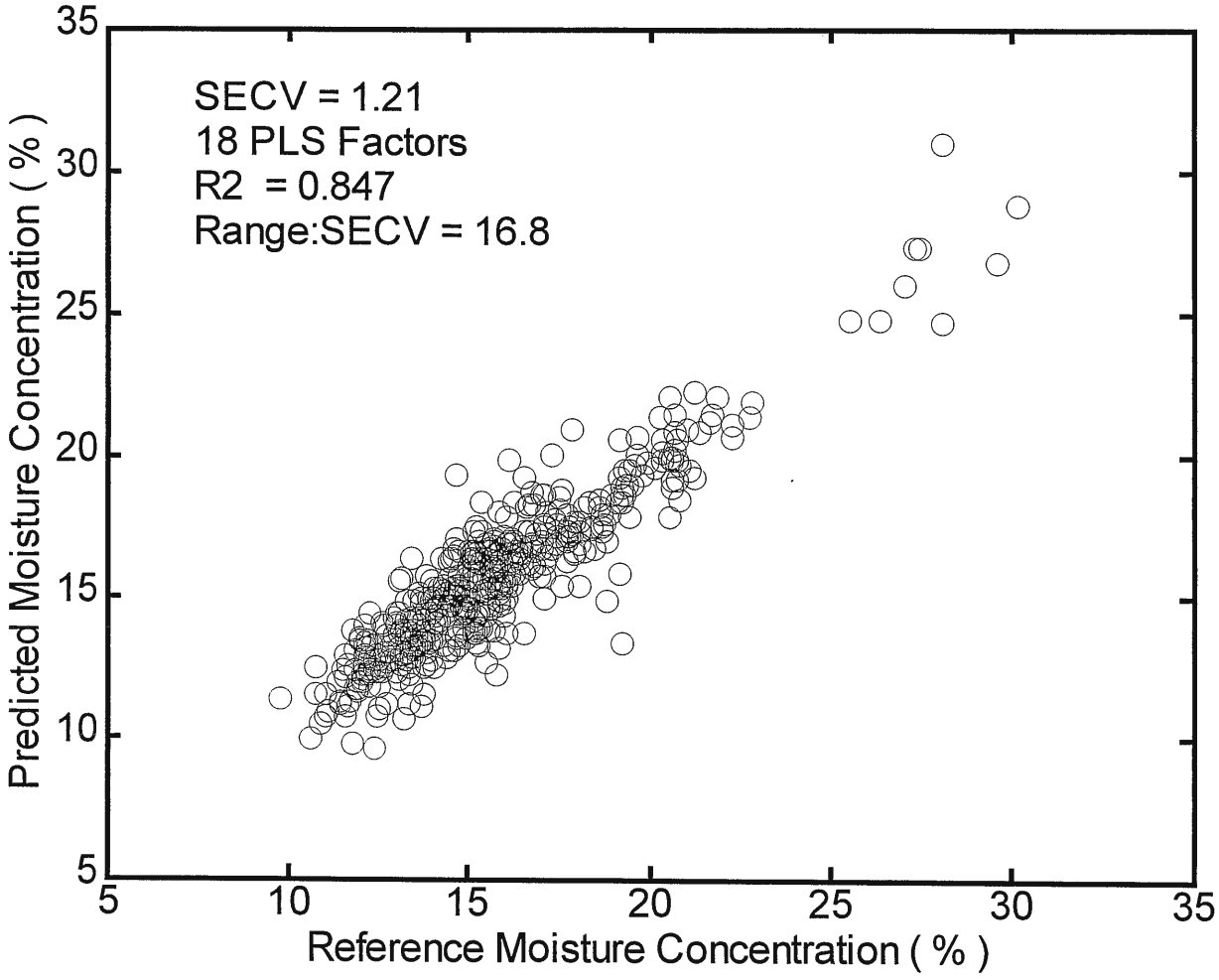

Figure 6. Summary of moisture calibration results before variable selection.

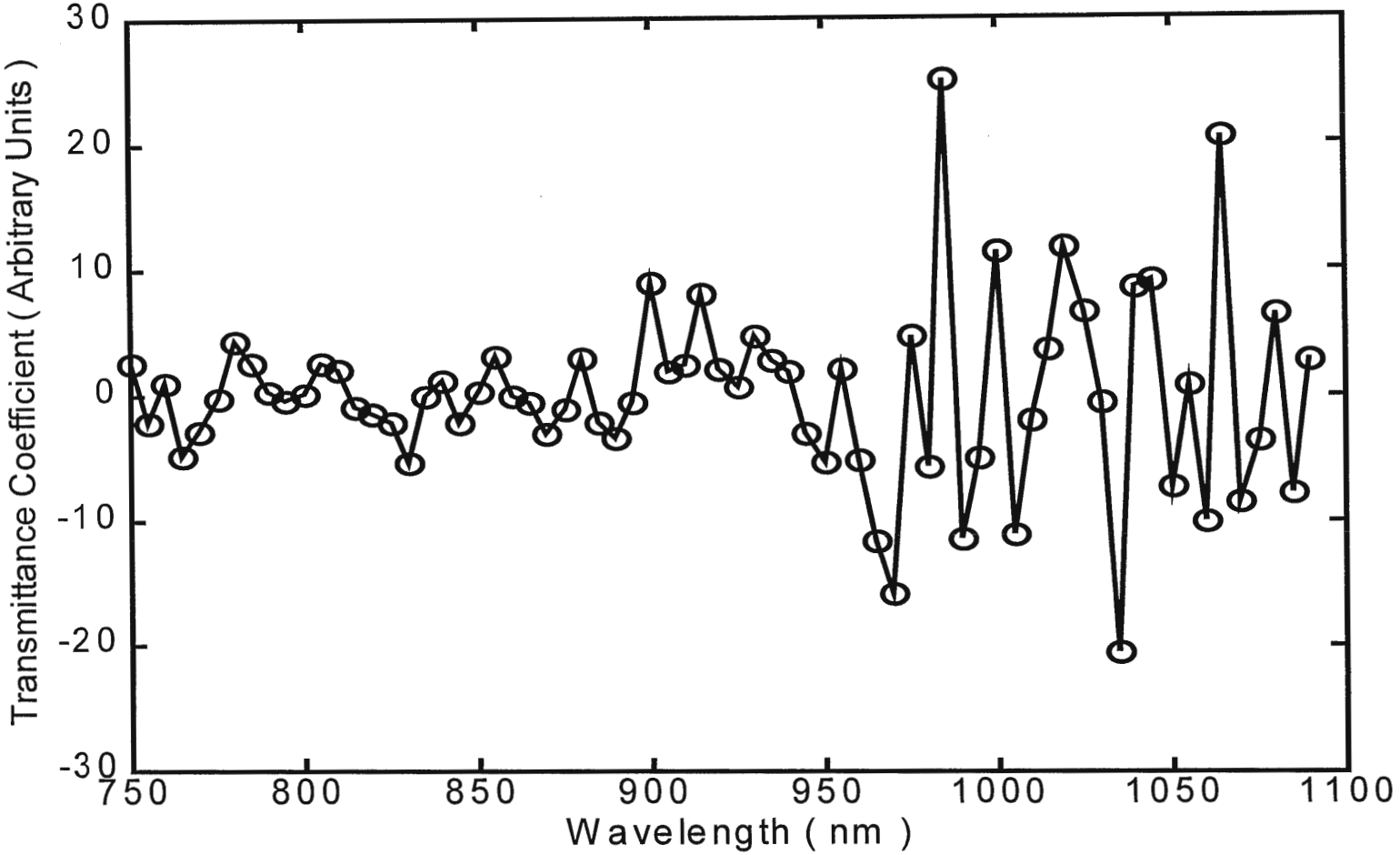

Figure 7. Moisture calibration regression coefficient vector.

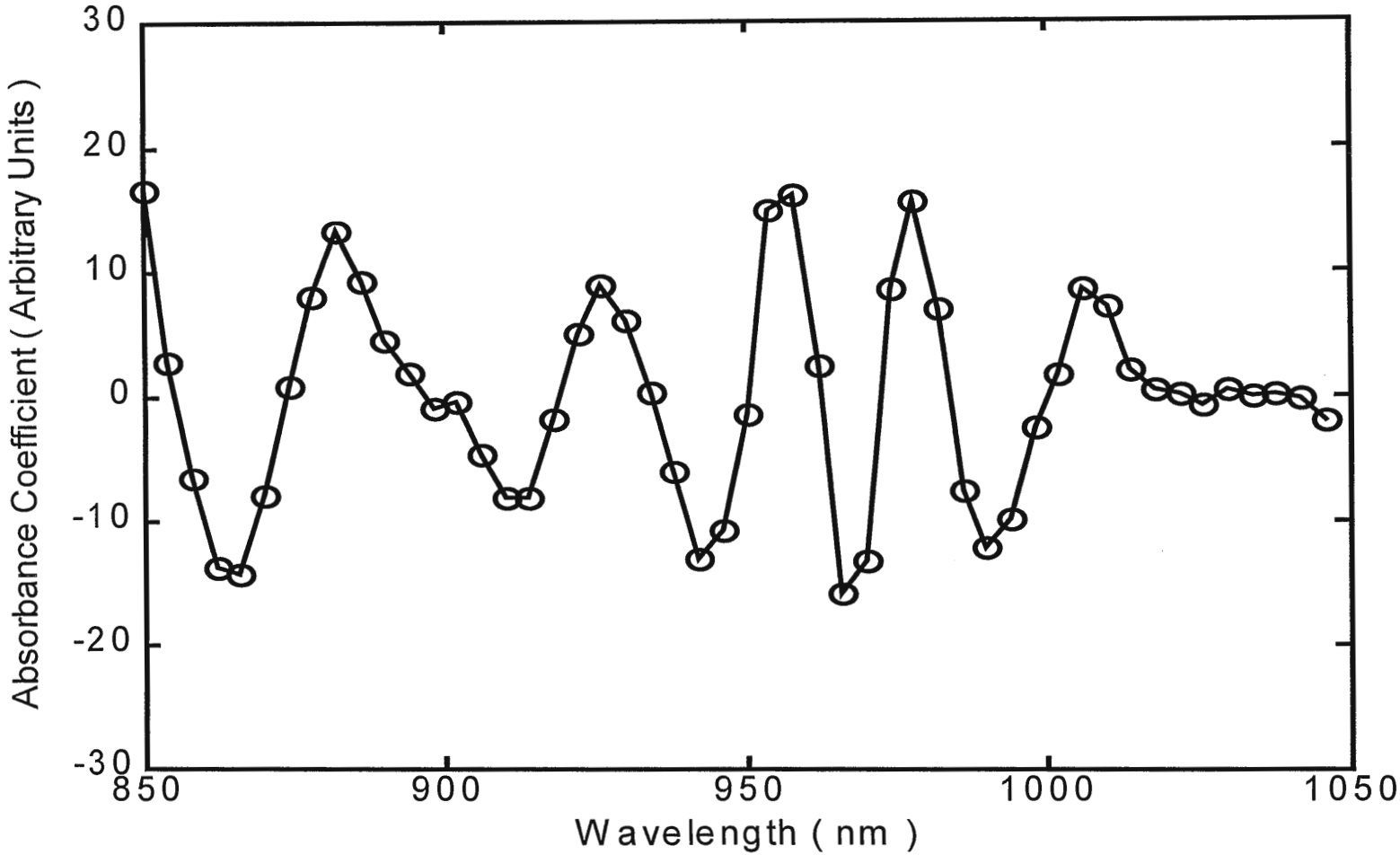

Figure 8. Infratec moisture calibration regression coefficient vector, derived using the same PLS algorithm as for the hyperspectral imaging spectrometer moisture calibration (4 nm wavelength spacing).

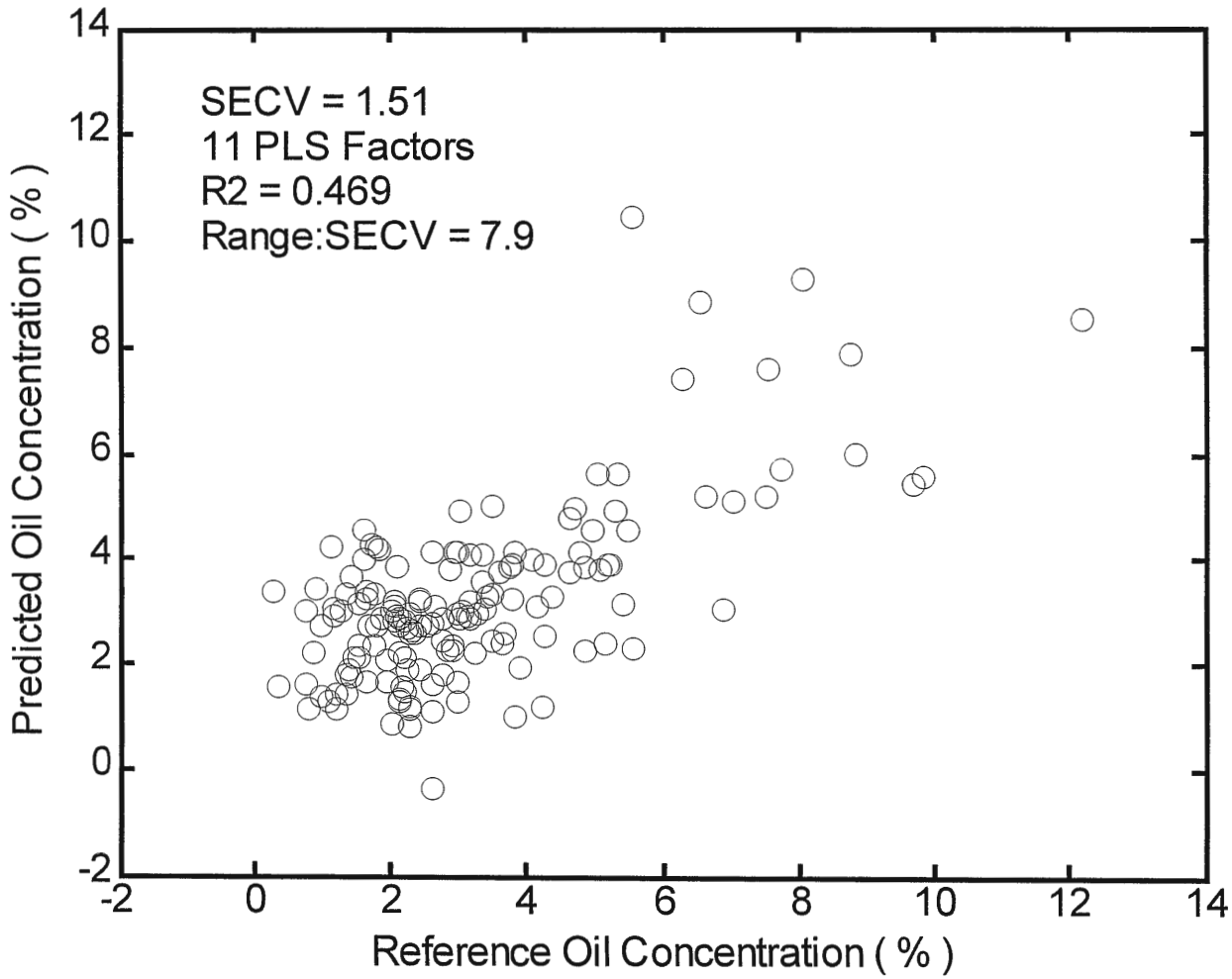

Figure 9. Summary of oil calibration results before variable selection.

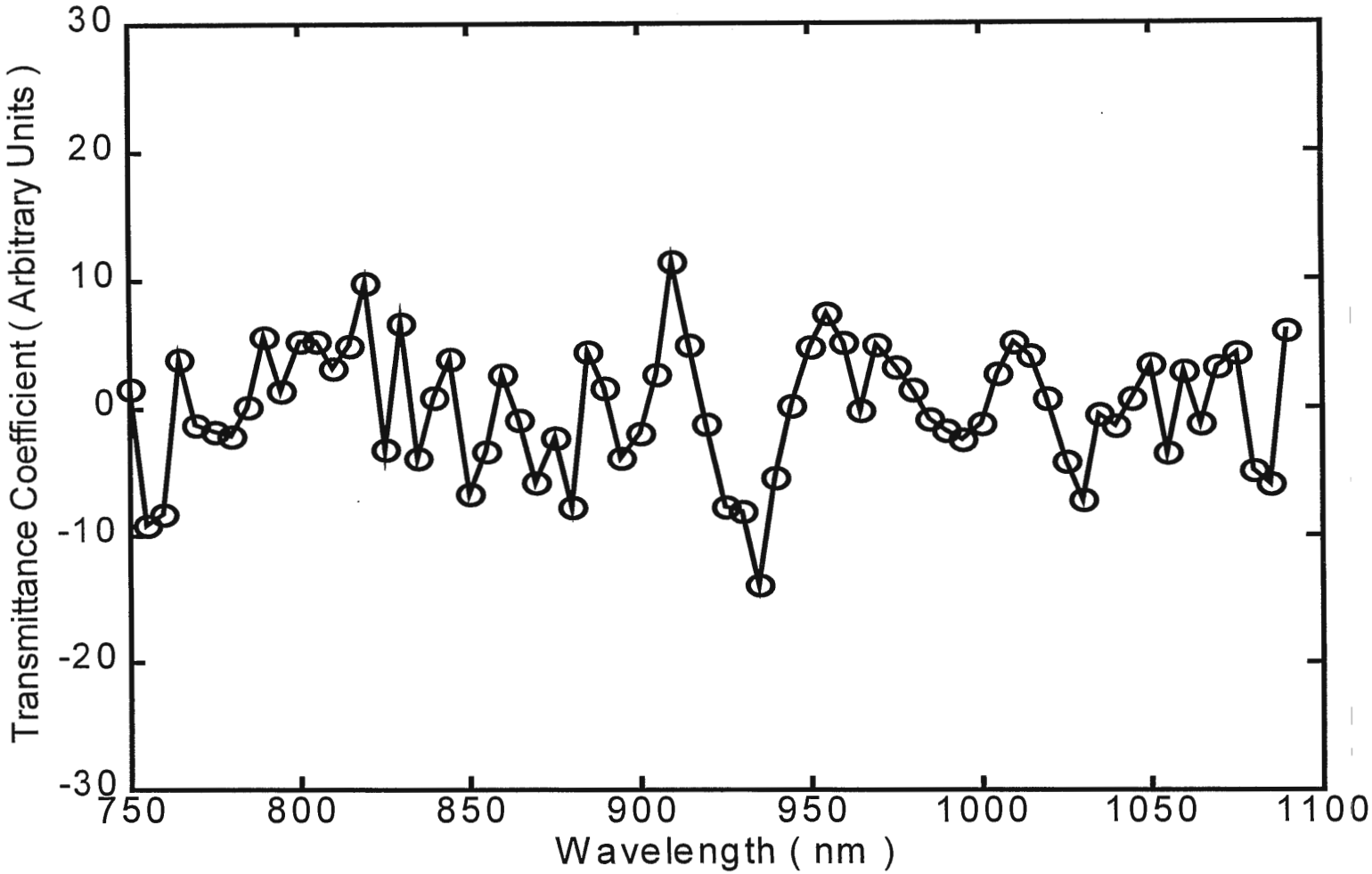

Figure 10. Oil calibration regression coefficient vector.

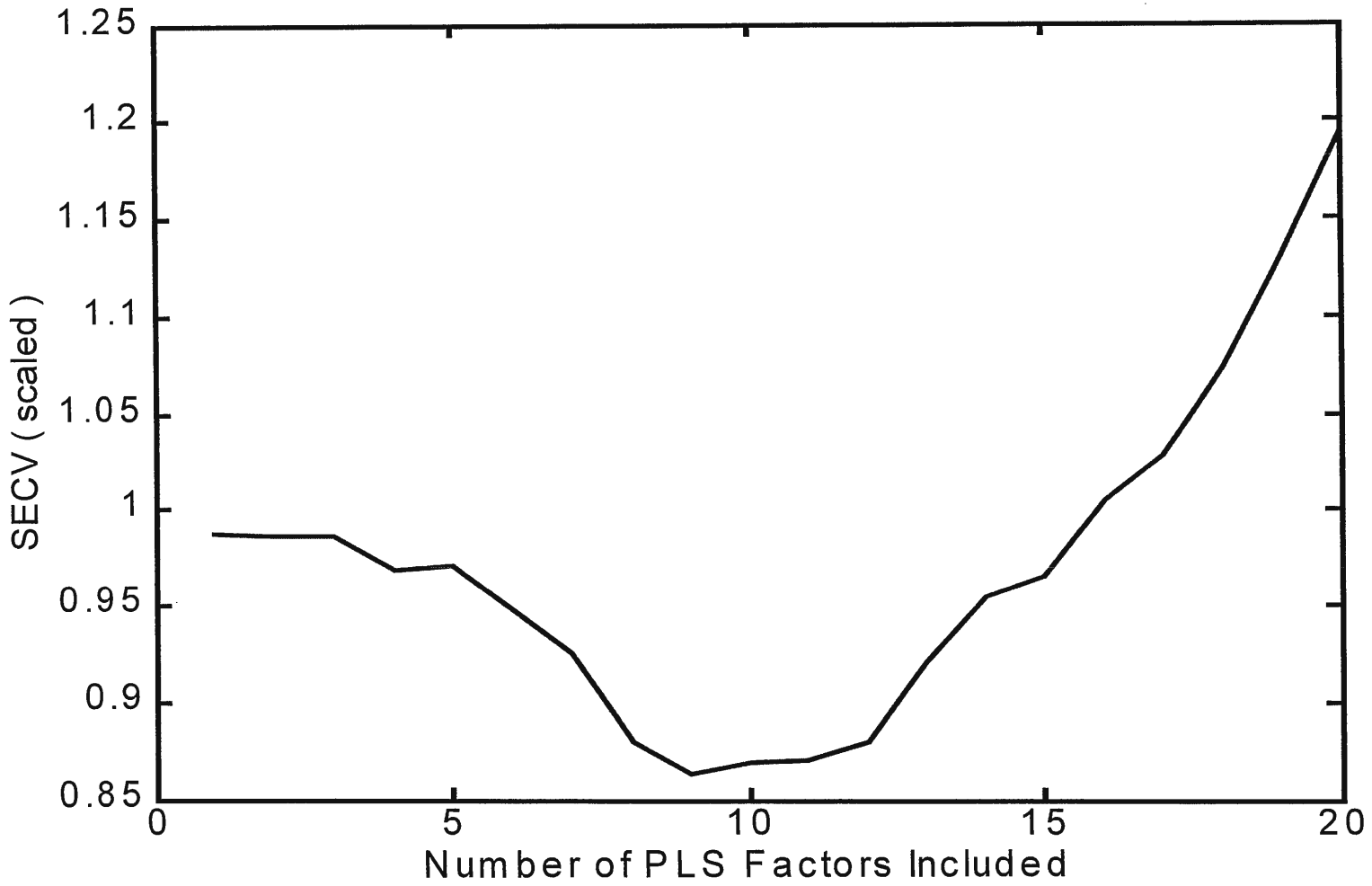

Figure 11. Scree plot of oil calibration PLS model development using cross-validation.

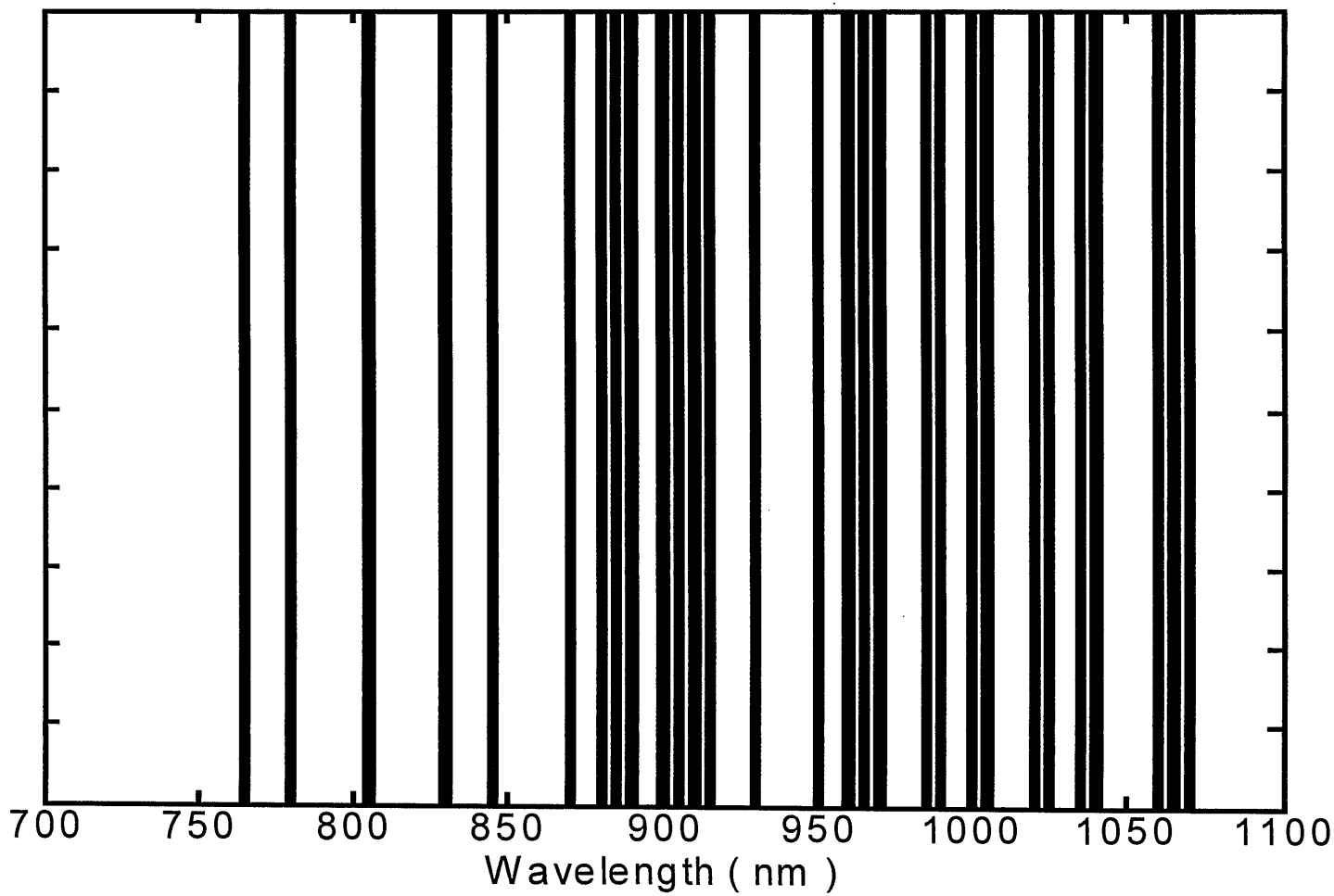

Figure 12. Wavelengths selected for optimal moisture calibration using a genetic algorithm and PLS with cross-validation.

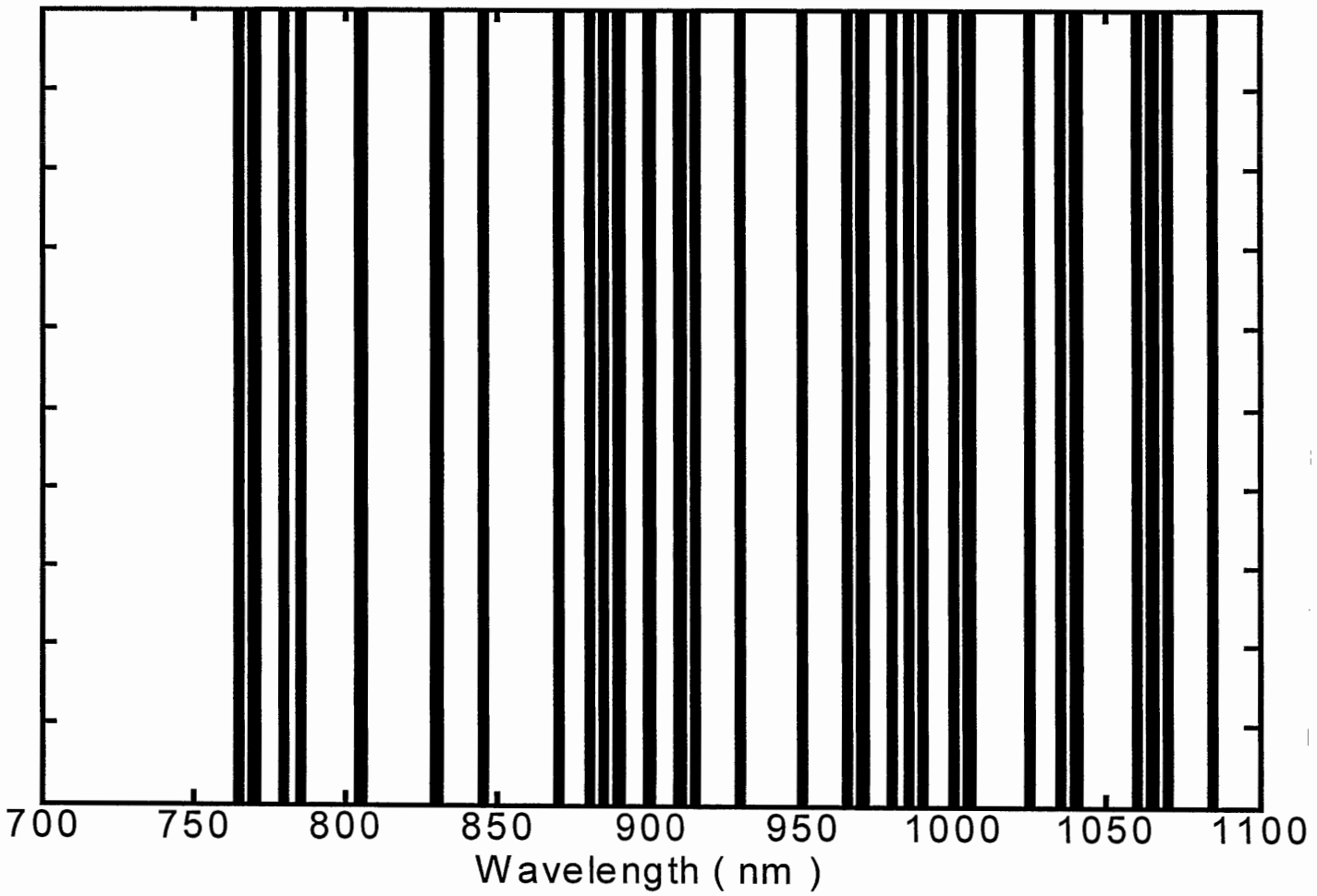

Figure 13. Wavelengths selected for optimal moisture calibration using a genetic algorithm and MLR with cross-validation.

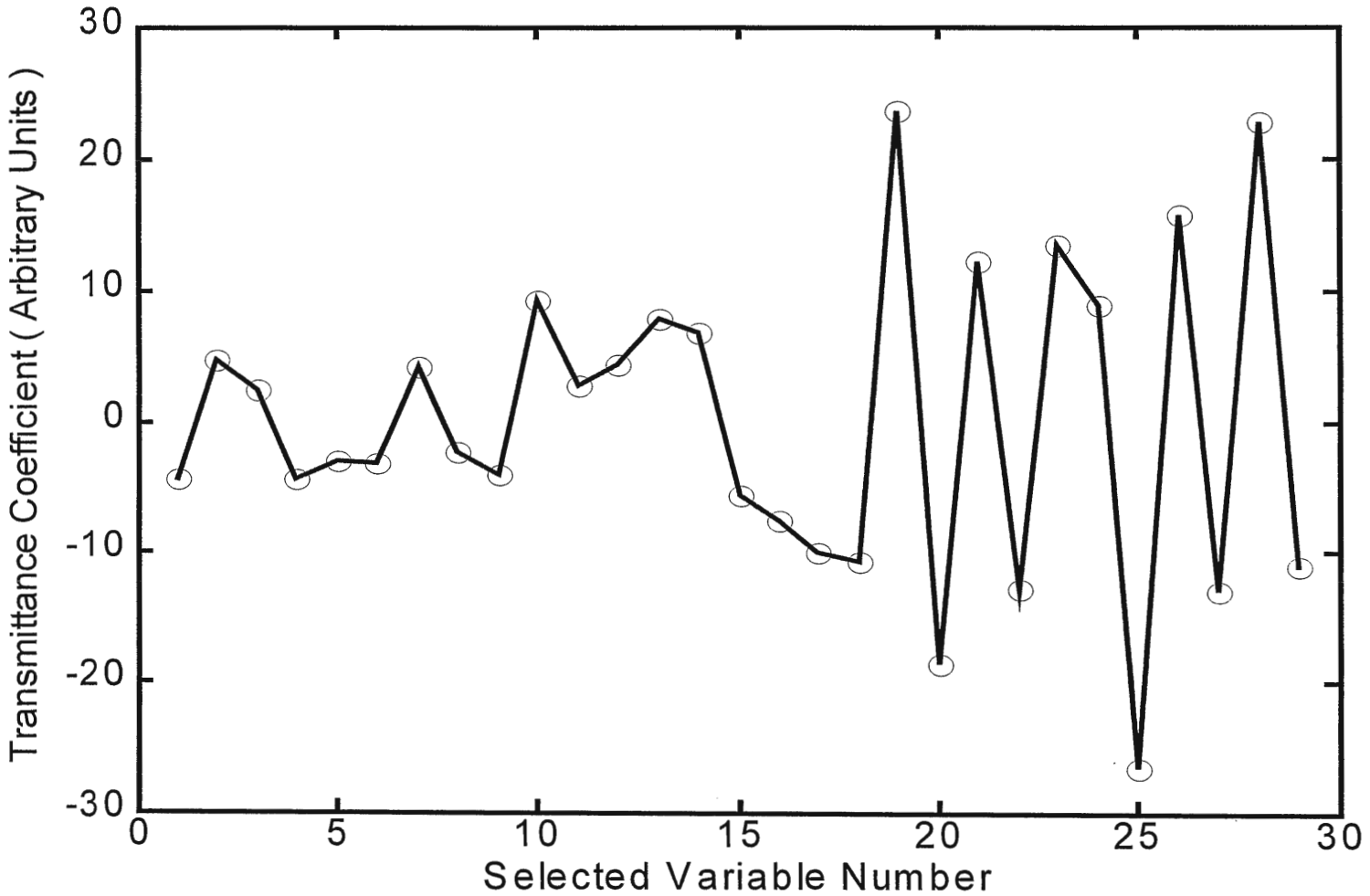

Figure 14. Moisture calibration coefficients, after variable selection by genetic algorithm.

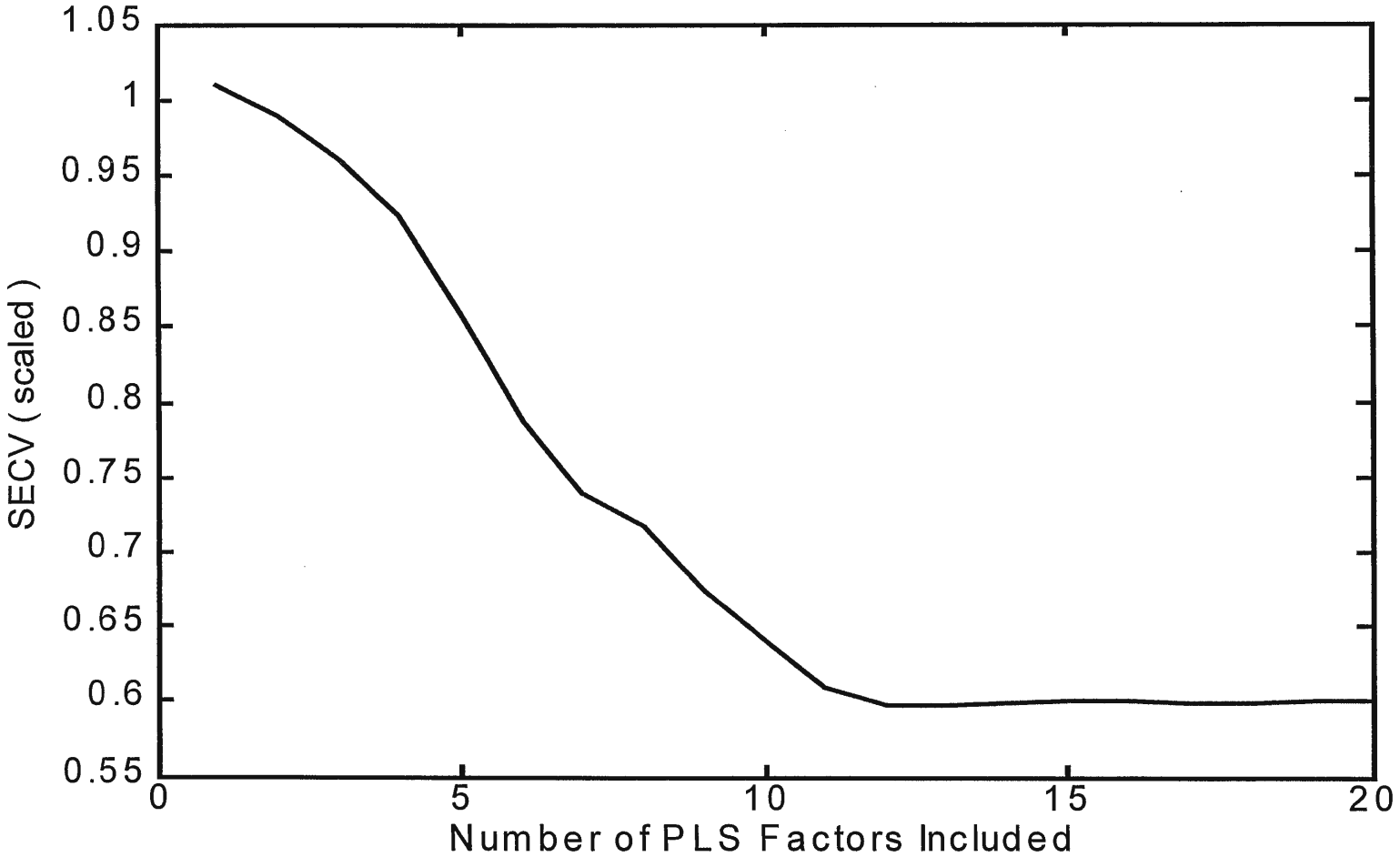

Figure 15. Scree plot of oil calibration PLS model development, using cross-validation, after variable selection by genetic algorithm.

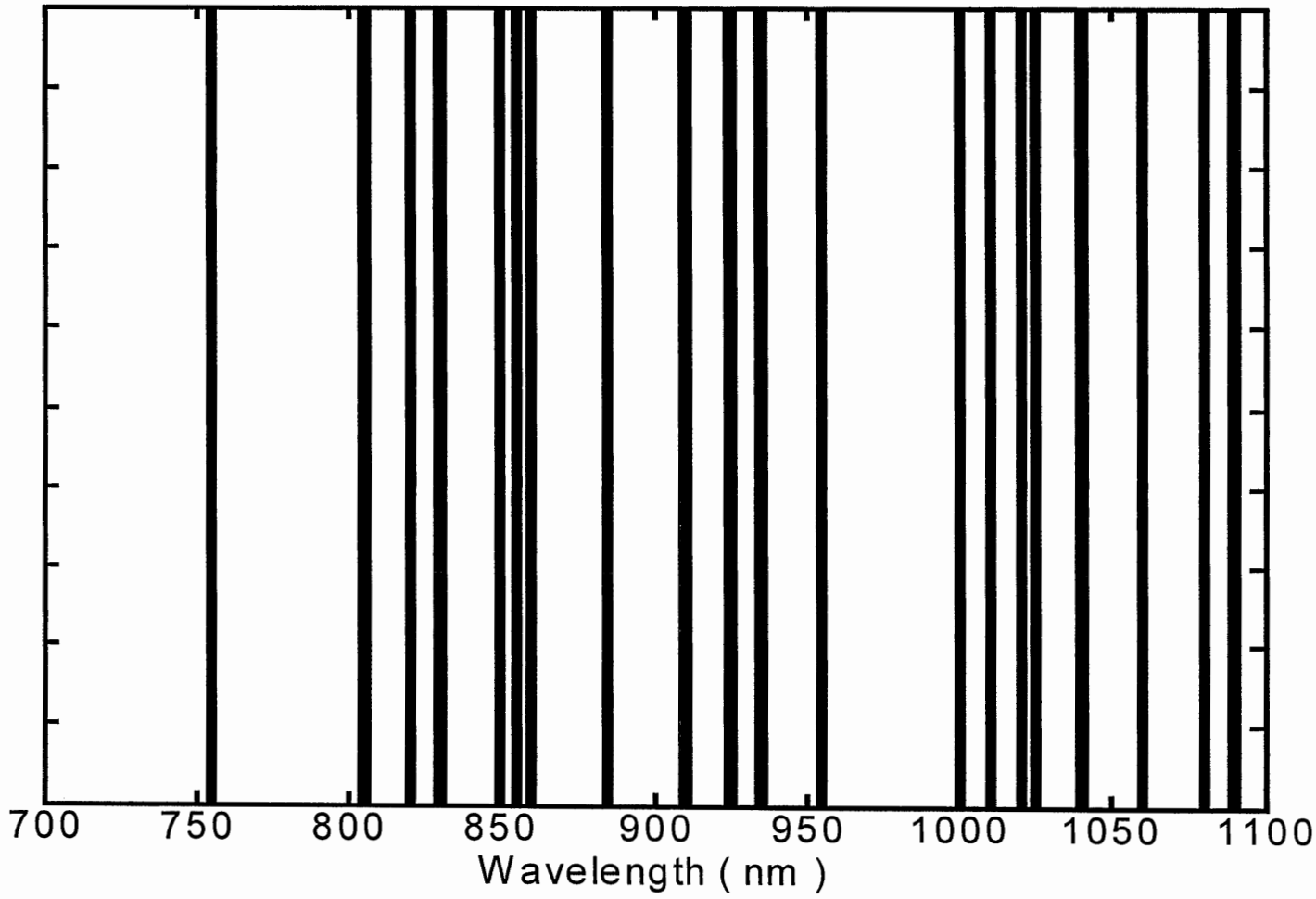

Figure 16. Wavelengths selected for optimal oil calibration using a genetic algorithm and PLS with cross-validation.

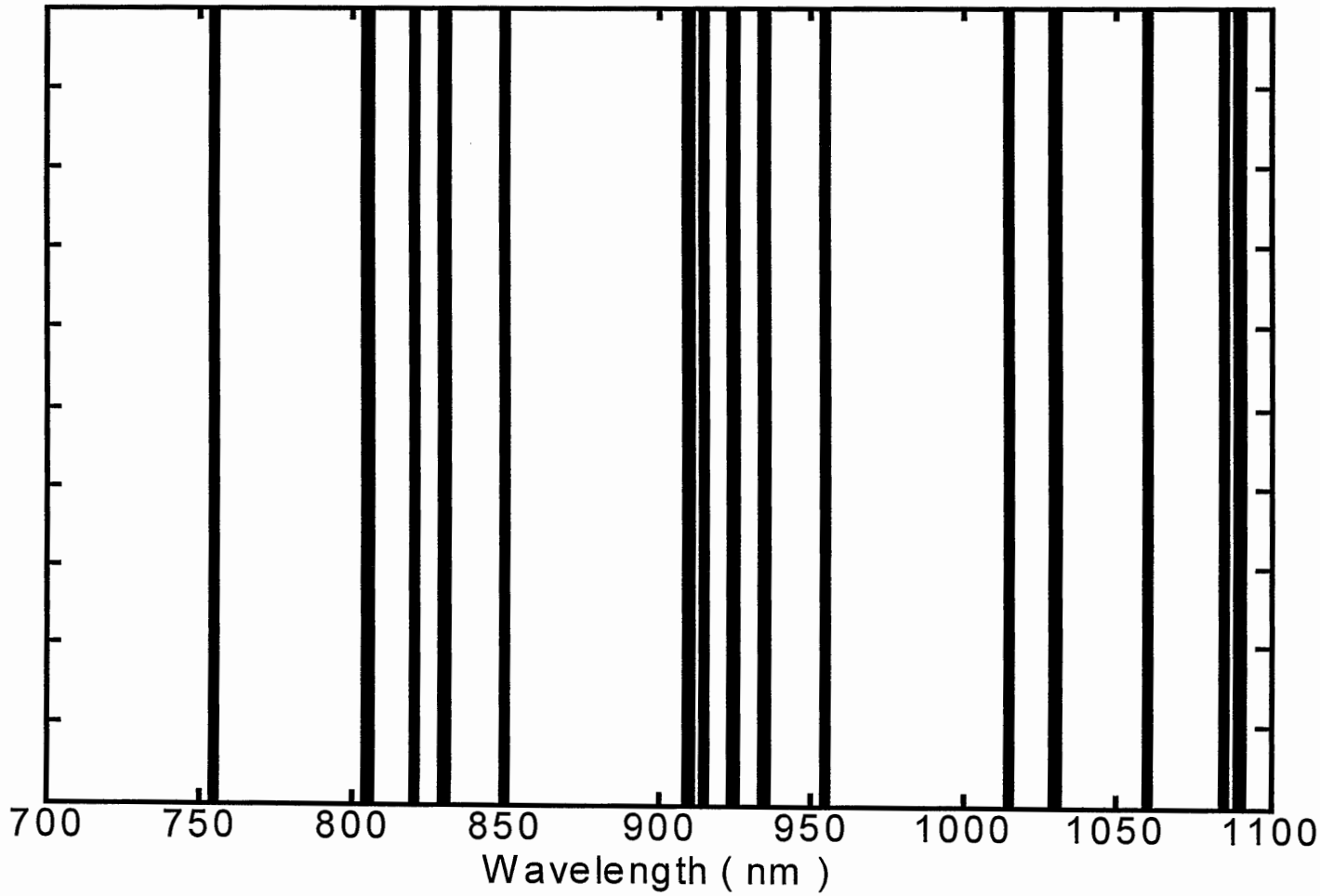

Figure 17. Wavelengths selected for optimal oil calibration using a genetic algorithm and MLR with cross-validation.

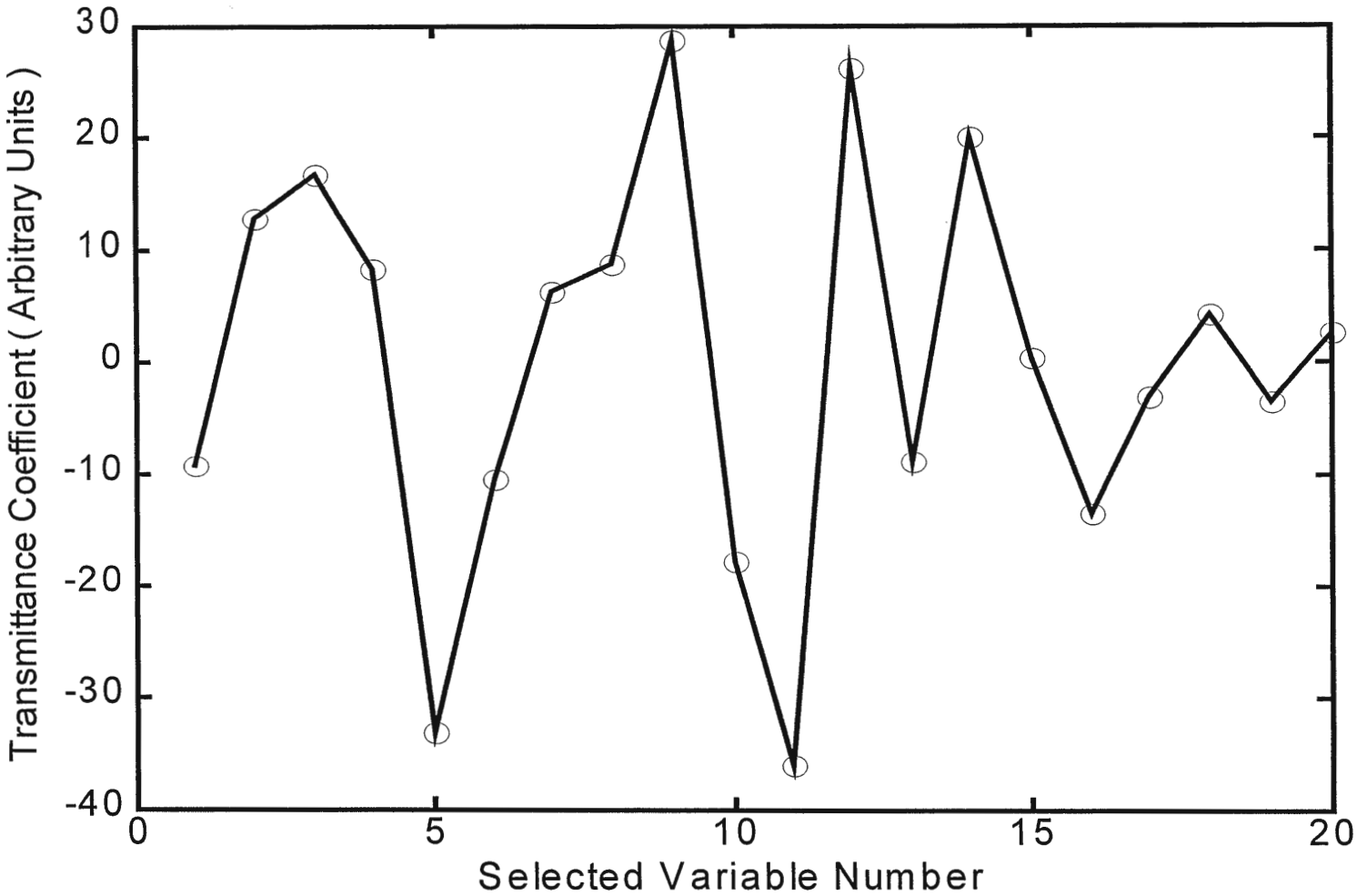

Figure 18. Oil calibration coefficients, after variable selection by genetic algorithm.

GENERAL CONCLUSIONS

A method of handling NIR hyperspectral image data was developed that maximized the cross-validation performance of the spectrometer in predicting the moisture content of single kernels of maize. Using the method developed and the current database of single-kernel maize NIR hyperspectral images and reference chemistry, oil and moisture calibrations can be developed that perform sufficiently well for segregation and quantitative quality prediction, respectively. In light of the information that this research has yielded, two clear directions emerge for the future development of single-kernel seed analysis by hyperspectral imaging.

In the first direction, the concept could be used to create a rapid quality analyzer, or a tool for segregating bulk grain samples into groups of single kernels of varying quality. In this scenario, beyond improving on the quality of the present calibrations, additional research would need to be conducted to improve the speed of hyperspectral image analysis. In this case, some further research might concentrate on:

1. Determining an optimal level of spatial resolution to maximize the accuracy versus speed tradeoff.
2. Investigating the effect of varying the level of A/D quantization accuracy.
3. Developing more efficient software routines to apply the algorithms used in this research.
4. Evaluating the long-term calibration stability of the NIR hyperspectral imaging spectrometer and the ability to transfer calibrations to other NIR imaging spectrometers.
5. Exploring other equipment designs to improve the speed and repeatability of single-seed NIR hyperspectral image analysis.

Alternatively, the results of this research might be applied to investigating the structure of single-seeds. While the focus of this research was on developing calibrations for predicting the average constituent concentration of a single kernel of maize, by modifying the techniques developed herein, it should be possible to develop suitable calibrations for predicting constituent concentrations at every pixel in a NIR

hyperspectral image of a single seed. This would allow visualizing spatial variability in composition. Moreover, seeds could be analyzed repeatedly in time to visualize the temporal changes in structure. In this case, beyond improving the speed and accuracy of the method, further research might concentrate on actions to:

1. Develop techniques for creating calibrations that are valid on a single-pixel basis within a NIR hyperspectral image.
2. Develop methods of modeling the physical structure of seeds using NIR hyperspectral images.
3. Investigate the use of classification tasks with NIR hyperspectral image datasets.
4. Develop methods for validating the performance of sub-kernel (i.e., single-pixel) calibrations.